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

PROCEEDINGS OF THE SIXTH INTERNATIONAL WORKSHOP ON MULTI-RELATIONAL DATA MINING

MRDM'07

September 17, 2007

Warsaw, Poland

Editors:
Donato Malerba
Department of Computer Science, University of Bari
Annalisa Appice
Department of Computer Science, University of Bari
Michelangelo Ceci
Department of Computer Science, University of Bari

Preface

The 6^{th} International Workshop on Multi-Relational Data Mining (MRDM 2007) was held in Warsaw, Poland, on September 17^{th} 2007 in conjunction with ECML/PKDD 2007: the 18^{th} European Conference on Machine Learning (ECML) and the 11^{th} European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD).

Data mining algorithms look for patterns in data. While most existing data mining approaches look for patterns in a single data table, multi-relational data mining (MRDM) approaches look for patterns that involve multiple tables (relations) from a relational database. Mining data which consists of complex/structured objects also falls within the scope of this field, since the normalized representation of such objects in a relational database requires multiple tables. Following the mainstream of MRDM research, the most common types of patterns and approaches considered in data mining have been extended to the multi-relational case and MRDM now encompasses relational association rule discovery, relational classification rules, relational decision and regression trees, and probabilistic relational models, among others. At same time, MRDM methods have been successfully applied across many application areas, ranging from the analysis of business data, through bioinformatics and pharmacology to Web mining and Spatial Data mining. Our goal is to bring together researchers and practitioners of data mining interested in methods for finding patterns in expressive languages from multi-relational / structured data and their applications. The workshop is the sixth of its kind. It follows the success of the workshops on Multi-Relational Data Mining, held both in Europe (ECML/PKDD 2001) and in USA (KDD 2002, 2003, 2004 and 2005).

Sixteen contributions were originally submitted, twelve of which were accepted for presentation. Each submission was evaluated by three independent referees. Besides paper presentations, the scientific programme also featured an invited talk by Luc De Raedt (Department of Computer Science, Katholieke Universiteit Leuven, Belgium).

We would like to thank the invited speaker, all the authors who submitted papers and all the workshop participants. We are also grateful to members of the program committee members and external referees for their thorough work in reviewing submitted contributions with expertise and patience. A special thank is due to both the ECML/PKDD Workshop Chair and the members of ECML/PKDD Organizing Committee who made this event possible.

Warsaw, September 2007

Donato Malerba
Annalisa Appice
Michelangelo Ceci

Workshop Organization

Workshop Chairs

Donato Malerba	University of Bari - Italy
Annalisa Appice	University of Bari - Italy
Michelangelo Ceci	University of Bari - Italy

Program Committee

Hendrik Blockeel	Katholieke Universiteit Leuven - Belgium
Jean-François Boulicaut	INSA Lyon - France
Sašo Džeroski	Jožef Stefan Institute - Slovenia
Peter Flach	University of Bristol - UK
Thomas Gärtner	Fraunhofer Institute for Autonomous Intelligent Systems - Germany
Lise Getoor	University of Maryland - USA
David Jensen	University of Massachusetts - USA
Kristian Kersting	MIT Computer Science and Artificial Intelligence Laboratory - USA
Joerg-Uwe Kietz	Kdlabs AG, Zurich - Switzerland
Arno Knobbe	Universiteit Utrecht - The Netherlands
Joost Kok	Leiden University - The Netherlands
Stefan Kramer	Technical University Munich - Germany
Nada Lavrač	Jožef Stefan Institute - Slovenia
Celine Rouveirol	University Paris Sud XI - France
Michele Sebag	University Paris Sud XI - France
Arno Siebes	Universiteit Utrecht - The Netherlands
Stefan Wrobel	Fraunhofer Institute for Autonomous Intelligent Systems / University of Bonn - Germany

Additional Reviewers

Marenglen Biba	University of Bari - Italy
Anton Dries	Katholieke Universiteit Leuven - Belgium
Aneta Ivanovska	Jožef Stefan Institute - Slovenia
Arne Koopman	Universiteit Utrecht - The Netherlands
Christine Körner	Fraunhofer Institute for Autonomous Intelligent Systems - Germany
Wannes Meert	Katholieke Universiteit Leuven - Belgium
Jan Struyf	Katholieke Universiteit Leuven - Belgium
Bernard Zenko	Jožef Stefan Institute - Slovenia

Table of Contents

ProbLog and its Application to Link Mining in Biological Networks

(Invited Talk)

Luc De Raedt

Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200 A, B-3001 Heverlee, Belgium
luc.deraedt@cs.kuleuven.be

Abstract. ProbLog is a recently introduced probabilistic extension of Prolog [De Raedt, Kimmig, Toivonen, IJCAI 07]. A ProbLog program defines a distribution over logic programs by specifying for each clause the probability that it belongs to a randomly sampled program, and these probabilities are mutually independent. The semantics of ProbLog is then defined by the success probability of a query in a randomly sampled program. It has been applied to link mining and discovery in a large biological network. In the talk, I will also discuss various learning settings for ProbLog and link mining, in particular, I shall present techniques for probabilistic local pattern mining, probabilistic explanation based learning and theory compression from examples [De Raedt et al, ILP 96].

A Multi-Relational Approach to Clustering Trajectory Data

Gianni Costa, Alfredo Cuzzocrea, Giuseppe Manco,
Riccardo Ortale, and Howard Scordio

ICAR Inst., National Research Council, Italy
Via P. Bucci 41C, I-87036 Rende, Italy
{costa,cuzzocrea,manco,ortale,scordio}@icar.cnr.it

Abstract. We propose a novel methodology for clustering multi-relational trajectory data. Our methodology consists of two steps. Initially, tuple linkages, defined in the database schema of the multi-relational trajectories, are leveraged to *virtually* organize the available route data into as many transactions, i.e. as sets of feature-value pairs. The identified transactions are then partitioned into homogeneous groups. Each discovered cluster is equipped with a *representative*, that provides an explanation of the corresponding group of trajectories, in terms of those feature-value pairs that are most likely to appear in a transaction belonging to that particular group. Outliers trajectories are placed into a *trash cluster*, that is finally partitioned to mitigate the dissimilarity between the trash cluster and the previously generated clusters.

Full article in PDF

Choosing the Right Patterns: An Experimental Comparison between Different Tree Inclusion Relations

Jeroen De Knijf * and Ad Feelders

Algorithmic Data Analysis Group
Department of Information and Computing Sciences, Universiteit Utrecht
PO Box 80.089, 3508 TB Utrecht

Abstract. In recent years a variety of mining algorithms has been developed, to derive all frequent subtrees from a database of labeled ordered rooted trees. These algorithms share properties such as enumeration strategies and pruning techniques. They differ however in the tree inclusion relation used and how attribute values are dealt with. In this work we investigate the different approaches with respect to 'usefulness' of the derived patterns, in particular, the performance of classifiers that use the derived patterns as features. In order to find a good trade-off between expressiveness and runtime performance of the different approaches, we also take the complexity of the different classifiers into account, as well as the run time and memory usage of the different approaches. The experiments are performed on two real datasets. The results show that significant improvement in both predictive performance and computational efficiency can be gained by choosing the right tree mining approach.

Full article in PDF

* Supported by the Netherlands Organisation for Scientific Research (NWO) under grant no. 612.066.304.

Mining Frequent Patterns from Multi-Dimensional Relational Sequences

Nicola Di Mauro, Teresa M.A. Basile, Stefano Ferilli, and Floriana Esposito

Università degli Studi di Bari, Dipartimento di Informatica, 70125 Bari, Italy

Abstract. The problem addressed in this paper regards the discovering of frequent multi-dimensional patterns from relational sequences. In a multi-dimensional sequence each event depends on more than one dimension, such as in spatio-temporal sequences where an event may be spatially or temporally related to other events. In literature, the multi-relational data mining approach has been successfully applied to knowledge discovery from complex data. This work takes into account the possibility to mine complex patterns, expressed in a first-order language, in which events may occur along different dimensions. A complete framework and an *Inductive Logic Programming* algorithm to tackle this problem is presented with preliminary experiments focussing on artificial multi-dimensional sequences.

Full article in PDF

ILP: Compute Once, Reuse Often *

Nuno A. Fonseca[1], Ricardo Rocha[2], Rui Camacho[3], and Vítor Santos Costa[2]

[1] Instituto de Biologia Molecular e Celular (IBMC), Universidade do Porto, Portugal
nf@ibmc.up.pt
[2] DCC-FC, Universidade do Porto, Portugal
{ricroc,vsc}@ncc.up.pt
[3] Faculdade de Engenharia & LIAAD, Universidade do Porto, Portugal
rcamacho@fe.up.pt

Abstract. Inductive Logic Programming (ILP) is a powerful and well-developed abstraction for multi-relational data mining techniques. However, ILP systems are not particularly fast, most of their execution time is spent evaluating the hypotheses they construct. The evaluation time needed to assess the quality of each hypothesis depends mainly on the number of examples and the theorem proving effort required to determine if an example is entailed by the hypothesis. We propose a technique that reduces the theorem proving effort to a bare minimum and stores valuable information to compute the number of examples entailed by each hypothesis (using a tree data structure). The information is computed only once (pre-compiled) per example. Evaluation of hypotheses requires only basic and efficient operations on *trees*. This proposal avoids re-computation of hypothesis' value in theory-level search and cross-validation algorithms, whenever the same data set is used with different parameters. In an empirical evaluation the technique yielded considerable speedups.

Full article in PDF

* This work has been partially supported by Myddas (POSC/EIA/59154/2004) and by funds from the *Programa Operacional "Ciência, Tecnologia, Inovação" (POCTI) e do Programa Operacional "Sociedade da Informação" (POSI) do Quadro Comunitário de Apoio III (2000-2006)*. Nuno A. Fonseca is funded by FCT grant SFRH/BPD/26737/2005.

Mining Imbalanced Classes in Multirelational Classification

Hongyu Guo and Herna L. Viktor

School of Information Technology& Engineering,
University of Ottawa, Canada

Abstract. Multirelational classification algorithms search for patterns across multiple interlinked tables (relations) in a relational database. This type of method searches for relevant features both from a target relation (in which each tuple is associated with a class label) and relations related to the target, in order to better classify tuples in the target relation. Unfortunately, most of these methods implicitly assume that the classes in the target relation are equally represented. They thus tend to produce poor predictive performance over the underrepresented class in the data. This paper presents a novel strategy to deal with imbalanced multirelational data where the number of examples of one class in the target relation is much higher than the others. The algorithm learns from multiple views of a relational database and then combines the knowledge acquired by the view learners in order to find a better quality model for the skewed classes. Experiments performed on six real-world data sets show that the proposed method achieves promising results when compared with other popular relational data mining algorithms, in terms of the ROC curve and AUC value obtained. In particular, our method outperforms the others for very highly imbalanced data sets.

Stratified Gradient Boosting for Fast Training of Conditional Random Fields

Bernd Gutmann[1] and Kristian Kersting[2]

[1] Department of Computer Science, Katolieke Universiteit Leuven
Celestijnenlaan 200A, 3001 Heverlee, Belgium
[2] CSAIL, Massachusetts Institute of Technology
32 Vassar Street, Cambridge, MA, 02139-4307, USA

Abstract. Boosting has recently been shown to be a promising approach for training conditional random fields (CRFs) as it allows to efficiently induce conjunctive (even relational) features. The potentials are represented as weighted sums of regression trees that are induced using gradient tree boosting. Its large scale application such as in relational domains, however, suffers from two drawbacks: induced trees can spoil previous maximizations and the number of generated regression examples can become quite large. In this paper, we propose to tackle the latter problem by injecting randomness into the regression estimation procedure by subsampling regression examples. Experiments on a real-world data set show that this sampling approach is comparable with more sophisticated boosting algorithms in early iterations and, hence, provides an interesting alternative as it is much simpler to implement.

Full article in PDF

A Restart Strategy for Fast Subsumption Check and Coverage Estimation

Ondřej Kuželka and Filip Železný

Intelligent Data Analysis Research Group
Dept. of Cybernetics, Czech Technical University in Prague
http://ida.felk.cvut.cz
{kuzelo1,zelezny}@fel.cvut.cz

Abstract. We study the runtime distributions of a simple subsumption check algorithm and show that in some conditions they exhibit heavy tails, indicating a possible runtime advantage achievable by randomizing and restarting the algorithm. Therefore we design RESUMER, a restarted subsumption tester, incorporating randomization while preserving completeness. On generated graph data, RESUMER outperforms the state-of-the-art subsumption algorithm Django (i) significantly in the YES region of the phase transition domain and (ii) in the entire phase transition domain given a sufficient size difference between the tested subsumer and subsumee. Importantly, we further show how, under a distributional assumption, a restarted strategy can be used to quickly obtain a maximum likelihood estimate of the coverage of a pattern (proportion of examples subsumed thereby) without requiring to verify subsumption for all examples. We implement this technique in the program RECOVER and show that it provides accurate coverage estimates in favorable runtimes.

Full article in PDF

Relational Transformation-based Tagging for Human Activity Recognition

Niels Landwehr[1], Bernd Gutmann[1], Ingo Thon[1], Matthai Philipose[2], and Luc De Raedt[1]

[1] Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200 A, B-3001 Heverlee, Belgium
firstname.lastname@cs.kuleuven.be
[2] Intel Research Seattle
1100 NE 45th Street
Seattle, WA 98105, USA
matthai.philipose@intel.com

Abstract. The ability to recognize human activities from sensory information is essential for developing the next generation of smart devices. Many human activity recognition tasks are — from a machine learning perspective — quite similar to tagging tasks in natural language processing. Motivated by this similarity, we develop a relational transformation-based tagging system based on inductive logic programming principles, which is able to cope with expressive relational representations as well as a background theory. The approach is experimentally evaluated on two activity recognition tasks and compared to Hidden Markov Models, one of the most popular and successful approaches for tagging.

Full article in PDF

Learning Ground CP-logic Theories by means of Bayesian Network Techniques

Wannes Meert, Jan Struyf, and Hendrik Blockeel

Department of Computer Science, Katholieke Universiteit Leuven,
Celestijnenlaan 200A, 3001 Leuven, Belgium,
{wannes.meert, jan.struyf, hendrik.blockeel}@cs.kuleuven.be

Abstract. Causal relationships are present in many application domains. CP-logic is a probabilistic modeling language that is especially designed to express such relationships. This paper investigates the learning of CP-theories from examples, and focusses on structure learning. The proposed approach is based on a transformation between CP-logic theories and Bayesian networks, that is, the method applies Bayesian network learning techniques to learn a CP-theory in the form of an equivalent Bayesian network. We propose a constrained refinement operator for such networks that guarantees equivalence to a valid CP-theory. We experimentally compare our method to a standard method for learning Bayesian networks. This shows that CP-theories can be learned more efficiently than Bayesian networks given that causal relationships are present in the domain.

Full article in PDF

Learning Ground ProbLog Programs from Interpretations

Fabrizio Riguzzi

ENDIF, Università di Ferrara, Via Saragat 1, 44100 Ferrara, Italy
fabrizio.riguzzi@unife.it

Abstract. The relations between ProbLog and Logic Programs with Annotated Disjunctions imply that Boolean Bayesian networks can be represented as ground ProbLog programs and acyclic ground ProbLog programs can be represented as Boolean Bayesian networks. This provides a way of learning ground acyclic ProbLog programs from interpretations: first the interpretations are represented in tabular form, then a Bayesian network learning algorithm is applied and the learned network is translated into a ground ProbLog program. The program is then further analyzed in order to identify noisy or relations in it. The paper proposes an algorithm for such identification and presents an experimental analysis of its computational complexity.

Full article in PDF

Towards a Framework for Relational Learning and Propositionalization

Ulrich Rückert and Stefan Kramer

Institut für Informatik/I12, Technische Universität München, Boltzmannstr. 3, D-85748 Garching b. München, Germany,
{rueckert,kramer}@in.tum.de

Abstract. We present first steps towards a general framework for propositionalization and relational learning based on sequences of queries and models, and the information effectively needed to generate both. From an abstract point of view, we only consider sequences of queries sent to a database, and sequences of consecutive models that combine those queries in a decision function. On a more detailed and procedural level, we consider how the queries in a sequence are actually generated, and, in particular, which information is taken into account to do that. In this way, the framework can address the question of how well the provided information is used by different learning approaches. While we provide a categorization scheme for existing methods, the framework's main purpose is to address a number of theoretical and practical questions. On the theoretical side, questions concerning model selection and overfitting avoidance can be addressed. More practically, we present a simple visualization scheme comparing the generalization performance of methods. Finally, the framework could provide hints for software design or for combining known building blocks in novel ways.

Full article in PDF

Distributed Relational State Representations for Complex Stochastic Processes

Ingo Thon[1] and Kristian Kersting[2]

[1] Katholieke Universiteit Leuven, Department of Computer Science
Celistijnenlaan 200A, 3001 Heverlee, Belgium
ingo.thon@cs.kuleuven.be
[2] Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, 32 Vassar St, Cambridge, MA 02139, USA
kersting@csail.mit.edu

Abstract. Several promising variants of hidden Markov models (HMMs) have recently been developed to efficiently deal with large state and observation spaces and relational structure. Many application domains, however, have an apriori componential structure such as parts in musical scores. In this case, exact inference within relational HMMs still grows exponentially in the number of components. In this paper, we propose to approximate the complex joint relational HMM with a simpler, distributed one: k relational hidden chains over n states, one for each component. Then, we iteratively perform inference for each chain given fixed values for the other chains until convergence. Due to this structured mean field approximation, the effective size of the hidden state space collapses from $O(n^k)$ to $O(n)$.

Full article in PDF

Author Index