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

INTERNATIONAL WORKSHOP ON

CONSTRAINT-BASED MINING AND LEARNING

AT ECML/PKDD 2007

CMILE'07

September 21, 2007

Warsaw, Poland

Editors:
Siegfried Nijssen
Katholieke Universiteit Leuven
Luc De Raedt
Katholieke Universiteit Leuven

Preface

In many applications of machine learning and data mining the most interesting results are not obtained by a single run of a single algorithm. It is often necessary that a user be able to express constraints and preferences on the models and patterns that the algorithms have to output. This workshop intends to bring together researchers in data mining and machine learning that have an interest in mining and learning algorithms that explicitly offer users the possibility to express multiple types of constraints and preferences, and who believe that general ways to deal with —sometimes conflicting— constraints are necessary. Results of this kind have been obtained in constraint-based pattern mining algorithms, algorithms that learn decision trees under constraints and constraint-based clustering algorithms; also issues such as the language in which to express constraints, and approaches for dealing with multiple conflicting constraints, are important.

This workshop is the successor of the workshop on Knowledge Discovery in Inductive Databases (KDID). Constraint-based mining methods are a core component of Inductive Databases. The new name of the workshop reflects that we believe that constraints are not only important in data mining, but also in machine learning.

We decided to have four types of presentations. Besides invited presentations, we have presentations of abstracts and of full papers. In the abstract presentations work is presented that was previously published in other conferences, but which is of interest to the topic of the workshop. By including abstract presentations we hope to reduce the barrier for submitting high quality work to the workshop. In the paper presentations work is presented that was submitted to CMILE and reviewed by the program committee of the workshop. Depending on the enthousiasm of the program committee, these papers were accepted either as full presentations or as short presentations.

We look forward to an interesting workshop.

Leuven, August 2007

Siegfried Nijssen
Luc De Raedt

Workshop Organization

Workshop Chairs

Siegfried Nijssen (Katholieke Universiteit Leuven)
Luc De Raedt (Katholieke Universiteit Leuven)

ECML/PKDD Workshop Chair

Marzena Kryszkiewicz (Warsaw University of Technology)

Workshop Program Committee

Hiroki Arimura
Hendrik Blockeel
Francesco Bonchi
Jean-François Boulicaut
Toon Calders
Ian Davidson
Saso Dzeroski
Peter Flach
Minos Garofalakis
Thomas Gärtner
Fosca Giannotti
Bart Goethals
Jiawei Han
Tomer Hertz
Kristian Kersting
Ross D. King
Joost N. Kok
Stefan Kramer
Tilman Lange
Taneli Mielikäinen
Jan Struyf
Rosa Meo
Shinichi Morishita
Céline Robardet
Arno Siebes
Kiri Wagstaff
Takashi Washio
Philip S. Yu
Xifeng Yan
Mohammed Zaki

Table of Contents

Part I

Full Presentations – Abstracts

Anytime Learning of Cost-sensitive Decision Trees

Saher Esmeir and Shaul Markovitch

Computer Science Department, Technion-IIT, Isreal

Abstract. Machine learning techniques are increasingly being used to produce a wide-range of classifiers for complex real-world applications that involve different constraints both on the resources allocated for the learning process and on the resources used by the induced model for future classification. As the complexity of these applications grows, the management of these resources becomes a challenging task. In this work we introduce ACT (Anytime Cost-sensitive Tree learner), a novel framework for operating in such environments. ACT is an anytime algorithm that allows trading computation time for lower classification costs. It builds a tree top-down and exploits additional time resources to obtain better estimations for the utility of the different candidate splits. Using sampling techniques ACT approximates for each candidate split the utility of the subtree under it and favors a split with the best evaluation. Due to its stochastic nature ACT is expected to be able to escape local minima, into which greedy methods may be trapped. ACT can be applied in two anytime setups: the contract setup where the allocation of resources is known in advance, and the interruptible setup where the algorithm might be queried for a solution at any point of time. Experiments with a variety of datasets were conducted to compare the performance of ACT to that of the state of the art decision tree learners. The results show that for most domains ACT produces significantly better trees. In the cost-insensitive setup, where test costs are ignored, ACT could produce smaller and more accurate trees. When test costs are involved, the ACT trees were significantly more efficient for classification. ACT is also shown to exhibit good anytime behavior with diminishing returns.

S. Esmeir and S. Markovitch. Anytime Learning of Decision Trees. *Journal of Machine Learning Research (JMLR)*, 8, 2007.

S. Esmeir and S. Markovitch. Occam's Razor Just Got Sharper. In *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*, 2007.

S. Esmeir and S. Markovitch. Anytime Induction of Decision Trees: An Iterative Improvement Approach. In *Proceedings of The 21st National Conference on Artificial Intelligence (AAAI-2006)*, 2006.

S. Esmeir and S. Markovitch. When a Decision Tree Learner Has Plenty of Time. In *Proceedings of The 21st National Conference on Artificial Intelligence (AAAI-2006)*, 2006.

S. Esmeir and S. Markovitch. Lookahead-based Algorithms for Anytime Induction of Decision Trees. In *Proceedings of the 21st International Conference on Machine Learning (ICML-2004)*, 2004.

Constraint Based Hierarchical Clustering for Text Documents

Korinna Bade

University of Magdeburg, Germany

Abstract. This presentation deals with constraint based clustering in a hierarchical clustering scenario. The explicit goal is to derive a hierarchical structure of clusters that is constrained by a partially known hierarchy. In the presentation, related work in the field of constraint based clustering is analyzed, pointing out major drawbacks of the current approaches, especially when applied to a hierarchical scenario. Furthermore, special attention is drawn to problems concerning the clustering of text documents. Once problems are identified, two different types of approaches as well as their combination are presented that specifically target at finding a hierarchical cluster structure for text documents under constraints. Results of a first evaluation on a hierarchical dataset of text documents are shown, considering different aspects like unevenly distributed constraints. The presentation ends with proposals for future research in this area.

K. Bade and A. Nürnberger. Personalized hierarchical clustering. In: *Proceedings of the 2006 IEEE/WIC/ACM Int. Conference on Web Intelligence*, 2006.

The Chosen Few: On Identifying Valuable Patterns

Björn Bringmann and Albrecht Zimmermann

Katholieke Universiteit Leuven, Belgium

Abstract. Pattern search is one of the main topics in data mining. In the last years different types of pattern languages were developed in order to be able to deal with the ever new challenges of new and hopefully valuable representations for all kind of data. Most algorithms developed handle the usually computationally intensive task in a decent way enabling us to extract millions of patterns from even small datasets. These patterns are then presented to the user or they are used as features such that each instance of the original database can be converted into a binary vector, each bit encoding the presence or absence of the pattern.
Due to the fact that interestingness (i.e. constraint satisfaction) of patterns is evaluated for each pattern individually, the amount of patterns to be considered by a user is often too large. Furthermore, when presenting the binary vector data to a machine learning technique, an overabundance of features does not help in the learning task, possibly even "confusing" the algorithm, leading to overfitting.
The aim of our work is to select a small, human-interpretable subset con- taining little redundancy from a larger set of patterns while retaining as much information as possible encoded in the full set.
We define the information of a pattern set as the partition it induces on the database, with all instances sharing the exact same subset of patterns belonging uniquely to one block. Thus, information is obtained from the composition of all patterns, contrary to e.g. the notion of closed patterns. Closedness only considers patterns having different covers; thus two mutually exclusive patterns are closed but induce the same partition.
Since the high amount of patterns leads to an exponentially large search space of possible subsets, we present a heuristic technique to tackle the problem. The pattern set is processed in a given order and only patterns that effect a change to the current partition are selected. Besides arbitrary orderings, our rather general algorithm furthermore allows for different techniques for measuring the effect of the pattern, thus allowing intuitive descriptions of the selection step or even of properties for the final subset.
We give some rather intuitive examples for selection criteria and combine them with straight-forward ordering strategies. When evaluated on several UCI data sets the techniques reduce the set of closed patterns severely.

B. Bringmann and A. Zimmermann. The Chosen Few: On Identifying Valuable Patterns. In: *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)*, 2007.

Part II

Full Presentations – Papers

A Fast Algorithm for Mining Utility-Frequent Itemsets

Vid Podpečan[1], Nada Lavrač[1,2], and Igor Kononenko[3]

[1] Jozef Stefan Institute, Ljubljana, Slovenia
[2] University of Nova Gorica, Nova Gorica, Slovenia
[3] University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

Abstract. Utility-based data mining is a new research area interested in all types of utility factors in data mining processes and targeted at incorporating utility considerations in both predictive and descriptive data mining tasks. High utility itemset mining is a research area of utility-based descriptive data mining, aimed at finding itemsets that contribute most to the total utility. A specialized form of high utility itemset mining is utility-frequent itemset mining, which – in addition to subjectively defined utility – also takes into account itemset frequencies. This paper presents a novel efficient algorithm FUFM (Fast Utility-Frequent Mining) which finds all utility-frequent itemsets within the given utility and support constraints threshold. It is faster and simpler than the original 2P-UF algorithm (2 Phase Utility-Frequent), as it is based on efficient methods for frequent itemset mining. Experimental evaluation on artificial datasets show that, in contrast with 2P-UF, our algorithm can also be applied to mine large databases.

Full article in PDF

Mining Views: Database Views for Data Mining

Hendrik Blockeel[1], Toon Calders[2], Elisa Fromont[1], Bart Goethals[3], and Adriana Prado[3]

[1] Katholieke Universiteit Leuven, Belgium
[2] Technische Universiteit Eindhoven, The Netherlands
[3] Universiteit Antwerpen, Belgium

Abstract. We propose a relational database model towards the integration of data mining into relational database systems, based on the so called virtual mining views. We show that several types of patterns and models over the data, such as itemsets, association rules, decision trees and clusterings, can be represented and queried using a unifying framework. We describe an algorithm to push constraints from SQL queries into the specific mining algorithms. Several examples of possible queries on these mining views, using the standard SQL query language, show the usefulness and elegance of this approach.

Full article in PDF

Part III

Short Presentations – Papers

Onto4AR: A Framework for Mining Association Rules

Cláudia Antunes

Technical University of Lisbon, Portugal

Abstract. Despite the efforts made in last decades to center the process of knowledge discovery on the user, the balance between the discovery of unknown and interesting patterns is far from being reached. The discovery of association rules is a paradigmatic case, where this balance is quite difficult to establish. In this paper, we propose a new framework for pattern mining – the Onto4AR. This framework is centered on the use of ontologies, for the representation and introduction of domain knowledge into the mining process. By defining constraints based on an ontology, the framework provides a mining environment independent of the problem domain. With this simplification on the definition and use of constraints, the framework contributes to reduce the gap between discovered rules and user expectations.

Full article in PDF

Iterative Constraints in Support Vector Classification with Uncertain Information

Jianqiang Yang and Steve Gunn

University of Southampton, UK

Abstract. This paper proposes a new iterative approach of input uncertainty classification, which incorporates input uncertainty information and exploits adaptive constraints extracted from uncertain inputs statistically and geometrically to extend the traditional support vector classification (SVC). Kernel functions can be implemented by a novel kernelized formulation to generalize this proposed technique to non-linear models and the resulting optimization problem is a second order cone program (SOCP) with a unique solution. Results demonstrate how this technique has an improved performance and is more robust than the traditional algorithms when uncertain information is available.

Full article in PDF

Multitarget Polynomial Regression

Aleksandar Pečkov, Sašo Džeroski, and Ljupčo Todorovski

Jozef Stefan Institute, Slovenia

Abstract. The paper addresses the task of multi-target polynomial regression, i.e., the task of inducing polynomials that can predict the value of more then one numeric variable. As in other learning tasks, we face the problem of finding an optimal trade-off between the complexity of the induced model and its predictive error. We propose a minimal description length scheme for multi-target polynomial regression, which includes coding schemes for polynomials and their predictive errors on training data. The proposed MDL scheme is implemented in an algorithm for polynomial induction that can also take into account language constraints, i.e., constraints on terms to be included in the induced polynomials. We empirically compare the multi-target model with the multiple single target models. The results of the experiments show that there is no loss in predictive performance when using multi-target models as compared to multiple target models and that fewer equation structures are considered in the former case.

Full article in PDF

Author Index