5th International Workshop on Knowledge Discovery in Inductive Databases (KDID'06): Workshop Report

Sašo Džeroski
Jožef Stefan Institute
Department of Knowledge Technologies
Jamova 39, 1000 Ljubljana, Slovenia
saso.dzeroski@ijs.si

Jan Struyf
Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium
jan.struyf@cs.kuleuven.be

ABSTRACT

This report presents a review of the 5th International Workshop on Knowledge Discovery in Inductive Databases (KDID'06), which was organized by the authors and held in Berlin, Germany, on September 18, 2006, in conjunction with ECML/PKDD'06, the 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. The goal of the workshop was to bring together the researchers that are interested in the area of inductive databases, inductive queries, constraint-based data mining, and data mining query languages.

1. INTRODUCTION

While knowledge discovery in databases (KDD) and data mining have enjoyed great popularity and success in recent years, there is a distinct lack of a generally accepted framework for data mining. The present lack of such a framework is perceived as an obstacle to the further development of the field (as discussed, e.g., during the SIGKDD'03 panel "Data Mining: The Next 10 Years"). The development of a unifying framework for data mining would clearly facilitate further progress in the field. The quest for such a framework is therefore a major research priority.

The most promising approach to this task is taken by inductive databases, an emerging research area at the intersection of data mining and databases. *Inductive databases (IDBs)* contain not only data, but also patterns. Patterns can be either *local patterns*, such as frequent itemsets, which are of descriptive nature, or *global models*, such as decision trees, which are of predictive nature. In an IDB, ordinary queries can be used to access and manipulate data, while *inductive queries* can be used to generate (mine), manipulate, and apply patterns. In the IDB framework, patterns become "first-class citizens" and KDD becomes an extended querying process in which both the data and the patterns that hold in the data are queried. IDB research thus aims at replacing the traditional KDD process model, where steps like pre-processing, data cleaning, and model construction follow each other in succession, by a simpler model in which all data pre-processing operations, data mining operations, as well as post-processing operations are queries to an inductive database and can be interleaved in many different ways.

The IDB framework is appealing as a theory for data mining, because it employs *declarative* queries instead of ad-hoc *procedural constructs*. As such, it holds the promise of facilitating the formulation of an "algebra" for data mining, the equivalent of Codd's relational algebra for databases. As declarative queries are often formulated using constraints, inductive querying is closely related to *constraint-based data mining* and is concerned with defining the necessary constraints and primitives for effective data mining. The IDB framework is also appealing for data mining applications, as it supports the whole KDD process. In inductive query languages, the results of one (inductive) query can be used as input for another: nontrivial multi-step KDD scenarios can be thus supported in IDBs, rather than just single data mining operations.

2. SUMMARY OF THE WORKSHOP

The aim of the workshop was to bring together the researchers that are interested in the area of inductive databases, inductive queries, constraint-based data mining, and data mining query languages.

This fifth edition of the workshop followed the previous four successful KDID workshops, all organized in conjunction with the European Conference on Machine Learning (ECML) and the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), over the period 2002 to 2005. KDID'02 was held in Helsinki, Finland, KDID'03 was held in Cavtat-Dubrovnik, Croatia, KDID'04 was held in Pisa, Italy, and KDID'05 was held in Porto, Portugal. The last two editions of the workshop have published proceedings in Springer's Lecture Notes in Computer Science Series and also the proceedings of this edition will be published by Springer (cf. references [3; 4; 5]).

Two European IST FET projects have been devoted to the topic of the workshop: the project cInQ "Consortium on Knowledge Discovery by Inductive Queries" (2001-2004) and the project IQ "Inductive Queries for Mining Patterns and Models" (2005-2008). KDID'06 was supported by IQ. More information is available on the IQ project website: `http://iq.ijs.si/`.

The workshop opened with an invited talk by Kiri L. Wagstaff from the Jet Propulsion Laboratory in Pasadena, California, USA. The topic of Kiri's talk was constrained clustering. Over the past five years, constrained (semi-supervised) clustering methods have become very popular,

motivated by applications such as gene clustering, document clustering, web search result clustering, and automatic lane finding from GPS traces. In her talk, Kiri identified three key open questions for the field of constrained clustering: how can the utility of a given constraint set be determined a-priori; how can the most useful constraints be actively solicited; and when should constraints be propagated or shared with neighboring points? Kiri stressed that addressing these questions is required before constrained clustering methods can be applied to very large data sets in an efficient and principled fashion.

The workshop continued with the presentation of eleven contributed papers. Six of these considered local (frequent) pattern mining, two focused on global models, and three contributions were about inductive query languages and environments. We briefly discuss each contribution below.

Two contributions introduce new pattern representations. In "Quantitative Episode Trees", Mirco Nanni and Christophe Rigotti introduce a new pattern domain called quantitative episode trees for mining (sets of) event sequences. A quantitative episode tree compactly represents the main occurrence groups of an episode by means of a tree structure, and provides quantitative bounds on the durations of each step of the episode.

The second new representation is Zero-suppressed Binary Decision Diagrams (ZBDDs). In their paper "Frequent Pattern Mining and Knowledge Indexing Based on Zero-suppressed BDDs", Shin-ichi Minato and Hiroki Arimura propose to use ZBDDs to represent sets of item sets. ZBDDs generalize BDDs, which are a compact, graph based representation for Boolean functions. The paper proposes a (maximal) item set mining algorithm based on ZBDDs and argues that the ZBDD representation allows for efficient inductive querying of the item sets.

The next two contributions consider new constraint types. In "Efficient Mining under Flexible Constraints through Several Datasets", Arnaud Soulet, Jiří Kléma and Bruno Crémilleux propose a pattern mining algorithm supporting constraints built from a large set of primitives, which take into account heterogeneous data (e.g., binary data and texts). It relies for efficiency on a new closure operator that allows for interval based pruning of the search space. The algorithm is tested on bioinformatic data.

The second special type of constraints that is considered are so called soft constraints. In their work "Weighted and Probabilistic Instances of the Soft Constraint Based Pattern Mining Paradigm", Stefano Bistarelli and Francesco Bonchi provide a theoretical basis and experimental analysis of probabilistic and weighted soft constraints. The problem with regular constraints, such as minimum frequency, is that they are Boolean: a discontinuity occurs where the constraint changes from true to false. Interestingness, however, is usually not a discontinuous function and soft constraints address this mismatch.

Most constrained based pattern mining algorithms consider binary data. Practical data, however, may be numeric or might contain missing values. The following two contributions address these issues. In "Mining Bi-sets in Numerical Data", Jérémy Besson, Céline Robardet, Luc De Raedt and Jean-François Boulicaut propose a method for mining numerical bi-sets. A numerical bi-set is a set of objects and a set of attributes such that all values in the corresponding cells are within a user specified range. The latter can be viewed as a constraint on the bi-sets. A second constraint ensures that the bi-sets are maximal. The algorithm uses these constraints to prune the search for bi-sets.

Data with missing values is considered by François Rioult and Bruno Crémilleux in "Mining Correct Properties in Incomplete Databases" for the case of k-free patterns. The authors propose a new definition for k-freeness suitable for incomplete data. This definition guarantees that the k-free patterns extracted from an incomplete database are k-correct, that is, they are also k-free in every possible completion of the database. They also show how patterns satisfying this new definition can be efficiently mined.

Two papers investigate how inductive databases can better support global models, such as decision trees. Élisa Fromont and Hendrik Blockeel focus on this topic in "Integrating Decision Tree Learning into Inductive Databases". Inspired by a similar approach for item sets, they investigate how decision trees can be stored in relational tables and queried using standard SQL queries. Using a prototype implementation, they illustrate the method with a number of interesting inductive queries for decision trees.

Decision trees can be used for a wide range of data mining tasks, such as prediction and clustering, and are as such very relevant to IDBs (which should provide a general data mining framework). This is argued by Sašo Džeroski, Ivica Slavkov, Valentin Gjorgjioski and Jan Struyf in "Analysis of Time Series Data with Predictive Clustering Trees". This paper shows how predictive clustering trees, a generalization of decision trees, can cluster bioinformatic time series data. The main advantage of the approach over regular clustering is that it, in addition to the clustering, also provides a symbolic description of the clusters.

The last three papers consider inductive query languages and environments. In "IQL: A Proposal for an Inductive Query Language", Siegfried Nijssen and Luc De Raedt introduce the inductive query language IQL. IQL intends to be a general, descriptive, declarative, extendable, and implementable language that supports the mining of both local and global patterns, reasoning about inductive queries and query processing using logic, as well as the flexible incorporation of new primitives and solvers. IQL extends the tuple relational calculus with functions, types, and data mining primitives.

Kenneth A. Kaufman, Ryszard S. Michalski, Jarosław Pietrzykowski and Janusz Wojtusiak present their VINLEN system in "An Integrated Multi-task Inductive Database and Decision Support System VINLEN: An Initial Implementation and First Results". VINLEN is built around knowledge generation operators, which given input data and/or knowledge create new knowledge. The central operator of VINLEN is a natural induction module that generates hypotheses from data in the form of interpretable attributional rules. The paper illustrates VINLEN with a medical application.

Last but not least, Francesco Bonchi, Fosca Giannotti, Claudio Lucchese, Salvatore Orlando, Raffaele Perego and Roberto Trasarti present their ConQueSt system in "On Interactive Pattern Mining from Relational Databases". ConQueSt supports the intrinsically exploratory, human-guided, interactive, and iterative nature of pattern discovery. Following the IDB vision, it provides users with an expressive constraint based query language that allows the discovery process to be effectively driven toward potentially

interesting patterns. ConQueSt is a practical system that can mine "real world" data stored in relational databases.

3. CONCLUSION

We presented a brief review of the 5th International Workshop on Knowledge Discovery in Inductive Databases (KDID'06), which continued the tradition of the four previous editions and aimed to bring together the researchers that are interested in the area of inductive databases, inductive queries, constraint-based data mining, and data mining query languages. We certainly hope that the momentum gained by this 5th edition of the KDID workshop will continue to foster close cooperation between all researchers interested in these related fields.

This paper includes only a brief summary of the interesting contributions presented at KDID'06. The reader is therefore encouraged to consult the workshop web site at `http://www.cs.kuleuven.be/~dtai/KDID06` for additional information. The workshop notes, including the full text of all papers, can be obtained from the workshop page of the ECML/PKDD'06 website at `http://www.ecmlpkdd2006.org/workshops.html`. Extended versions of the workshop papers will soon appear in a volume dedicated to the workshop of Springer's Lecture Notes in Computer Science Series.

Acknowledgments

KDID'06 was supported by the European project IQ ("Inductive Queries for Mining Patterns and Models", IST FET FP6-516169, 2005-2008). Jan Struyf is a post-doctoral fellow of the Fund for Scientific Research of Flanders, Belgium (FWO-Vlaanderen).

4. REFERENCES

[1] L. De Raedt, F. Gianotti, R. Meo, and M. Klemettinen, editors. *The 1st International Workshop on Knowledge Discovery in Inductive Databases*, August 2002. Helsinki, Finland. `http://ecmlpkdd.cs.helsinki.fi/kdid-2002.html`.

[2] J.-F. Boulicaut and S. Džeroski, editors. *The 2nd International Workshop on Knowledge Discovery in Inductive Databases*, September 2003. Cavtat-Dubrovnik, Croatia. `http://www.cinq-project.org/ecmlpkdd2003`.

[3] B. Goethals and A. Siebes, editors. *Knowledge Discovery in Inductive Databases, 3rd International Workshop, KDID 2004, Pisa, Italy, September 20, 2004, Revised Selected and Invited Papers*, volume 3377 of *Lecture Notes in Computer Science*. Springer, 2005. `http://kdid04.cs.helsinki.fi`.

[4] F. Bonchi and J.-F. Boulicaut, editors. *Knowledge Discovery in Inductive Databases, 4th International Workshop, KDID 2005, Porto, Portugal, October 3, 2005, Revised Selected and Invited Papers*, volume 3933 of *Lecture Notes in Computer Science*. Springer, 2006. `http://www-kdd.isti.cnr.it/kdid05`.

[5] S. Džeroski and J. Struyf, editors. *Knowledge Discovery in Inductive Databases, 5th International Workshop, KDID 2006, Berlin, Germany, September 18, 2006, Revised Selected and Invited Papers*, 2006. To appear in Springer's Lecture Notes in Computer Science Series. `http://www.cs.kuleuven.be/~dtai/KDID06`.