Multi-Relational Data Mining: a Workshop Report

Sašo Džeroski
Dept. of Intelligent Systems
Jožef Stefan Institute
Jamova 39
SI-1000 Ljubljana
Slovenia
saso.dzeroski@ijs.si

Luc De Raedt
Machine Learning Lab
Institut fuer Informatik
Albert-Ludwigs-University Freiburg
Georges Koehler Allee 79
D-79110 Freiburg, Germany
deraedt@informatik.uni-freiburg.de

ABSTRACT

In this report, we briefly review the Multi-Relational Data Mining workshop, which was held in Edmonton, Canada on July, 23, 2002 as part of the workshop program of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-02).

1. INTRODUCTION

An increasing number of data mining applications involve the analysis of complex and structured types of data (such as sequences in genome analysis, HTML and XML documents) and require the use of expressive pattern languages. Many of these applications cannot be solved using traditional data mining algorithms. This observation forms the main motivation for the multi-disciplinary field of Multi-Relational Data Mining (MRDM). The solution proposed by MRDM consists of dealing with data in the form of multiple tables in a relational database. Mining data that represent complex/structured objects also falls within the scope of this field, since the normalized representation of such objects in a relational database requires the use of multiple tables.

The field of MRDM aims at integrating results from existing fields such as inductive logic programming, KDD, machine learning and relational databases; producing new techniques for mining multi-relational data; and practical applications of such techniques. Typical data mining approaches look for patterns in a single relation of a database. For many applications, squeezing data from multiple relations into a single table requires much thought and effort and can lead to loss of information. An alternative for these applications is to use multi-relational data mining. Multi-relational data mining can analyze data from a multi-relation database directly, without the need to transfer the data into a single table first. Thus the relations mined can reside in a relational or deductive database. Using multi-relational data mining it is often also possible to take into account background knowledge, which can be regarded as views in the database.

Present MRDM approaches consider all of the main data mining tasks, including association analysis, classification, clustering, learning probabilistic models and regression. The pattern languages used by single-table data mining approaches for these data mining tasks have been extended to the multiple-table case. Relational pattern languages now include relational association rules, relational classification rules, relational decision trees, and probabilistic relational models, among others. MRDM algorithms have been developed to mine for patterns expressed in relational pattern languages. Typically, data mining algorithms have been upgraded from the single-table case: for example, distance-based algorithms for prediction and clustering have been upgraded by defining distance measures between examples/instances represented in relational logic.

MRDM methods have been successfully applied across many application areas, ranging from the analysis of business data, through bioinformatics (including the analysis of complete genomes) and pharmacology (drug design) to Web mining (information extraction from text and Web sources).

The aim of the workshop was to bring together researchers and practitioners of data mining interested in methods for finding patterns in expressive languages from complex / multi-relational / structured data and their applications.

2. OVERVIEW OF THE CONTRIBUTIONS

The nine contributions presented at the workshop can be clustered around the following topics: graph- and sequence based approaches, multi-relational classification, and probabilistic multi-relational representations.

2.1 Graph- and Sequence Based Data

Analyzing graph- and sequence data has been investigated for quite some time in the field of multi-relational data mining. Graph theory is well-understood today. Furthermore, graphs can easily and elegantly represent structured objects, which explains the relation to the workshop theme.

A first contribution *Concept Formation Using Graph Grammars* (I. Jonyer, L. B. Holder, and D. J. Cook) investigated the use of graph-grammars for concept-formation and general machine learning or data mining tasks. Graph-grammars are a generalization of the usual type of grammars in that they do not only deal with sequences but also with graphs. The authors demonstrate that graph-grammars are useful as a compact representation of graph data and also introduce some learning algorithms for inducing them.

In *Mining Patterns from Structured Data by Beam-wise Graph-Based Induction* (T. Matsuda, H. Motoda, T. Yoshida, and T. Washio) the authors report on improvements to their GBI (Graph-Based Induction) system, as well as a novel application of graph-based induction on a Hepatitis data set.

One of the central ideas in this work concerns the use of a canonical representation for graphs. Using a canonical form, the data mining algorithms can avoid generating identical patterns or graphs more than once.

Constraint-Based Mining of Sequences in SeqLog (S.D. Lee and L. De Raedt) addresses the mining of logical sequences within the inductive logic programming tradition. Logical sequences are sequences of logical atoms, such as *latex(kdd,tex), xdvi(kdd), dvips(kdd,dvi)*. The MineSeqLog algorithm is able to retrieve patterns of interest under a conjunction of anti-monotone and monotone constraints (e.g., a minimum frequency on the positives, and a maximum one on the negatives).

A final contribution related to this topic *Discovering Knowledge from Relational Data Extracted from Business News* (A. Bernstein, S. Clearwater, S. Hill, C. Perlich and F. Provost) was concerned with constructing graphs that summarize information about companies starting from newspaper articles. In contrast with the previous contributions, this work was not concerned with analyzing graph data, but rather with inferring knowledge in the form of graphs. Indeed, knowledge about companies is discovered from large corpora of newspaper articles. The authors clearly demonstrate that the extracted knowledge is useful for a wide range of tasks.

2.2 Probabilistic Relational Representations

An important stream within contemporary multi-relational data mining and inductive logic programming addresses the issue of combining relational (or first order) representations with probabilistic approaches.

The contribution *Schemas and Models* (D. Jensen and J. Neville) reviews various existing approaches to learning probabilistic relational representations and characterizes them in the Schema-Model framework. The framework identifies two key components: 1) a schema, which can be viewed as the underlying database schema of the representation, and 2) a model, which defines the probabilistic nature of the representation. After introducing the Schema-Model, the authors discuss how it is instantiated in specific probabilistic relational representations and use this to discuss future directions.

Statistical Models for Relational Data (L. Getoor, D. Koller, and B. Taskar) introduces a novel probabilistic relational representation, called SRM (Statistical Relational Model). An SRM is a statistical model of a particular database instantiation. It summarizes information about the frequencies of tuples as well as join operations. As such, SRMs can be employed for query approximation. The authors first introduce the semantics of SRMs, then discuss algorithms for estimating their parameters and learning their structure.

2.3 Multi-Relational Classification

Traditionally, multi-relational data mining has investigated how well-known data mining algorithms, such as decision tree learners, can be upgraded towards the use of multi-relational representations.

Experiments With MRDTL – A Multi-Relational Decision Tree Learning Algorithm (H. Leiva and V. Honavar) presents the multi-relational decision tree learner MRDTL and applies it to several benchmark data sets such as the KDD Cup 2001 data. MRDTL exploits SQL to learn directly from data in a relational database.

The paper *Hierarchical Multi-Classification* (H. Blockeel, M. Bruynooghe, S. Dzeroski, J. Ramon, and J, Struyf) focusses on the problem of prediction when each example is assigned not one class value, but rather several class values that belong to (possibly different levels of) a hierarchy. Such tasks are common, e.g., in functional genomics, where a gene can have more than one function and there are hierarchies of functions with several levels. The approach of predictive clustering trees is used to address the task at hand: experiments in several domains were presented.

In *Towards Structural Logistic Regression: Combining Relational and Statistical Learning* (A. Popescul, L.H. Ungar, S. Lawrence, and D.M. Pennock), various binary classification tasks from ResearchIndex (formerly Citeseer) are addressed with a combination of relational learning techniques with logistic regression. In this approach, the different documents and their relationships are first represented within a multi-relational representation. Secondly, interesting features are found using principles of propositionalization. Finally, these features are employed within a logistic regression module.

3. WEBSITE

An electronic version of the individual papers presented at the workshop, as well as the entire Workshop Notes is available from the Website of the workshop
`http://www-ai.ijs.si/SasoDzeroski/MRDM2002/`

4. PANEL OVERVIEW

The workshop concluded with a panel discussion titled *Multi-Relational Data Mining: The Way Ahead*. The panelists were: Hendrik Blockeel, Mark Craven, Luc De Raedt, Pedro Domingos, Sašo Džeroski, Lawrence Holder, David Jensen, and Hiroshi Motoda. Pedro Domingos stated that "The time is now ripe for RDM. Great potential lies in cosidering issues such as data streams and potential killer applications may be found in the areas of social networks and web mining." David Jensen discussed statistical issues in RDM, e.g., that samples are likely not be i.i.d. in RDM and what to do about it. Lawrence Holder and Hiroshi Motoda discussed general and specific issues concerning learning from graph representations. Other statements of the panelists include: "The different approaches to RDM should be integrated." (Hendrik Blockeel); "Explore combinations of design choices in RDM algorithms." (Mark Craven); "The future of RDM is in upgrading (probabilistic models, SVMs, neural networks) and downgrading (to make it more efficient for specific applications, e.g., on sequences). (Luc De Raedt)"; "Future research directions for RDM include inductive databases and integrating knowledge and data from different sources" (Sašo Džeroski).

5. CONCLUSIONS

As far as the organizers are concerned, the workshop succeeded in its goals. More precisely, the papers presented were representative of the current state-of-the-art in multi-relational data mining and this both in terms of topics and applications covered as well as representations employed. The topics included multi-relational classification, graph- and sequence based learning and probabilistic relational representations.

The applications addressed involved text and web mining, link analysis and evidence extraction, as well as bio- and chemo-informatics. The representations and underlying methodologies followed included those of relational learning, inductive logic programming, entity-relationship model, graph- and sequence based mining.

In addition to these scientific aspects, there was also an important new social aspect: it was the first time that a workshop on multi-relational data mining or inductive logic programming counted more American attendees than European and Japanese ones.

Finally, the authors hope that the interest in MRDM will continue and also, that the richness and diversity in approaches will remain as one of its distinguishing features. Therefore, the authors plan to organise further workshops on this topic as well as a special issue of the SIGKDD Explorations (to appear in 2003) on this topic.

About the Authors

Sašo Džeroski is a Senior Scientific Associate of the Department of Intelligent System, Jozef Stefan Institute, Ljubljana, Slovenia. His research interests include among others inductive logic programming (ILP) and relational data mining (RDM). He was involved in several international projects related to ILP and was the scientific coordinator of ILPnet2: The Network of Excellence in ILP. He was co-chair of the Seventh and Ninth International Workshops on ILP (ILP-97 and ILP-99) and co-chair of The Sixteenth International Conference on Machine Learning (ICML-99). He has also co-organized a number of events related to the topic of RDM, including the *Summer School on Relational Data Mining*, held in Helsinki in August 2002. He is the co-author/co-editor of three books in the areas of ILP/RDM: *Inductive Logic Programming: Techniques and Applications*, the first authored book on ILP; *Learning Language in Logic*, concerned with learning from natural language resources; and finally the book *Relational Data Mining*.

Luc De Raedt is presently a professor of computer science at the Albert-Ludwigs-University, Freiburg, Germany, where he chairs the Machine Learning Lab since 1999. Before moving to Freiburg, he was a part-time senior lecturer and post-doctoral researcher at the Katholieke Universiteit Leuven, Belgium, where he also obtained his Ph.D. thesis in 1991. He was a coordinator of the European ESPRIT projects on *Inductive Logic Programming* (1992-1999) and the key organizer of ECML-PKDD 2001. His current research interests lie in the areas of multi-relational data mining, inductive logic programming, constraint-based mining and inductive databases and their applications to bio- and chemo-informatics.

Workshop Organisers

- Sašo Džeroski, Jozef Stefan Institute, Slovenia
- Luc De Raedt, Albert-Ludwigs-University Freiburg, Germany
- Stefan Wrobel, Otto-von-Guericke University, Magdeburg, Germany

Program Committee

- Hendrik Blockeel (Katholieke Universiteit Leuven)
- Jean-Francois Boulicaut (University of Lyon)
- Diane Cook (University of Texas at Arlington)
- Mark Craven (University of Wisconsin at Madison)
- Luc Dehaspe (PharmaDM)
- Pedro Domingos (University of Washington)
- Peter Flach (University of Bristol)
- Lise Getoor (University of Maryland)
- David Jensen (University of Massachusets at Amherst)
- Ross King (University of Aberystwith)
- Stefan Kramer (Albert-Ludwigs-Universitaet Freiburg)
- Nada Lavrač (Jozef Stefan Institute)
- Donato Malerba (University of Bari)
- Heikki Mannila (Nokia Research/Helsinki Institute for Information Technology)
- Tom Mitchell (Carnegie Mellon University)
- Hiroshi Motoda (University of Osaka)
- Stephen Muggleton (Imperial College)
- David Page (University of Wisconsin at Madison)
- Foster Provost (Stern School of Business, New York University)
- Celine Rouveirol (University Paris Sud XI)
- Gunter Saake (Otto-von-Guericke Universitaet Magdeburg)
- Michele Sebag (University Paris Sud XI)
- Arno Siebes (Universiteit Utrecht)
- Hannu Toivonen (University of Helsinki / Nokia Research)

Acknowledgements

The authors are grateful to Stefan Wrobel, who co-organized the workshop, to Osmar Zaiane, who helped with local organization, to the authors of the submitted papers, and the participants of the MRDM workshop.