Multi-Relational Data Mining 2004: Workshop Report

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

Hendrik Blockeel
Katholieke Universiteit Leuven
Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium
hendrik.blockeel@cs.kuleuven.ac.be

ABSTRACT

In this report we briefly review the 3rd Workshop on Multi-Relational Data Mining (MRDM-2004), which was organized by the authors and held in Seattle, WA, on August 22, as part of the workshop program of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The goal of the workshop was to bring together researchers and practitioners of Data Mining and interested in methods and applications of finding patterns in expressive languages from multi-relational, complex, and/or structured data.

1. INTRODUCTION

Relational data mining (RDM) approaches (Džeroski and Lavrač 2001) look for patterns that involve multiple tables (relations) from a relational database. To emphasize this fact, RDM is often referred to as multi-relational data mining (MRDM). Mining data which consists of complex/structured objects falls within the scope of MRDM, since the normalized representation of such objects in a relational database requires multiple tables. MRDM is a multi-disciplinary field, which 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. MRDM-2004 was the third edition of this Workshop on Multi-Relational Data Mining.

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 often corresponds to 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 accross 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).

2. SUMMARY OF THE WORKSHOP

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.

This workshop was the third of its kind. It followed the success of the first and second workshop on Multi-Relational Data Mining, held at SIGKDD 2002 and 2003, reports on which appear in SIGKDD Explorations [Vols 4(2) and 5(2)]. Further information on those workshops can be found at web sites MRDM-2002 or MRDM-2003. Based on MRDM-02, a special issue of SIGKDD Explorations [Vol 5(1)] was co-edited by Sašo Džeroski and Luc de Raedt.

In total, some 30 to 40 people attended the workshop, which consisted of two invited presentations and five contributed papers. We briefly summarize them here.

The workshop opened with an invited talk by Lise Getoor, from the University of Maryland, on link mining. Link mining considers the task of analysing data where individuals are linked to other individuals. It is a relatively new but fast-growing field, the interest in which is motivated by application areas such as bio-informatics, web mining, analysis of bibliographic citations, social network analysis, etc. In her talk, Lise Getoor presented an overview of the field, discussing among other things the different types of task that can be distinguished, such as link type prediction (what kind of relationship holds between two objects), link existence prediction (which objects are linked), object identification (discovering that different objects actually refer to the same individual), etc. A number of practical applications were discussed as an illustration.

The program continued with three regular contributions. The first speaker, Parag, presented *Multi-Relational Record*

Linkage, authored by Parag and Pedro Domingos. The authors consider the task of determining when two records actually refer to the same object (for instance, in a bibliography database, records with slightly different titles may actually refer to the same paper). They argue that making this decision independently for each pair of records is not optimal; when two records are matched, this may be evidence for certain related objects (e.g., attribute values) to match as well, and this should be taken into account when matching other record pairs. In general, evidence for matches should propagate to related objects. The authors propose a multi-relational method based on conditional random fields that does exactly that, and experimentally show its merit.

The next speaker was Jesse Davis, presenting work on *Using Bayesian Classifiers to Combine Rules*, by Jesse Davis, Vítor Santos Costa, Irene Ong, David Page, and Inês Dutra. This work is set in the context of inductive logic programming. ILP systems typically learn a set of rules that are combined in a disjunctive way (an example is positive if at least one rule fires). This may lead to many false positives when rules are not entirely accurate. The authors discuss how combining the rules in a more sophisticated way, more specifically using bayesian networks, alleviates this problem; moreover, bayesian nets can cater for dependencies among the rules, which further improves performance.

Finally, Daan Fierens presented *Logical Bayesian Networks*, by Daan Fierens, Hendrik Blockeel, Maurice Bruynooghe and Jan Ramon. Logical Bayesian Networks are yet another formalism for describing probabilistic logical models. Fierens started by discussing some strengths and weaknesses of two existing frameworks, Probabilistic Relational Models and Bayesian Logic Programs, in terms of expressiveness and interpretability. He next proposed Logical Bayesian Networks as a novel framework that has been designed specifically to combine the strengths of both.

In the afternoon, Jiawei Han, from the University of Illinois at Urbana-Champaign, gave the second invited talk, titled "CrossMine: Efficient Classification Across Multiple Database Relations". The CrossMine approach that he discussed bears a strong resemblance to classical inductive logic programming (ILP) techniques, but is set in the framework of relational databases. ILP techniques are known to be quite inefficient, and the CrossMine approach contains two important innovations that alleviate this problem. The first is the use of tuple ID propagation, which can be seen as a method to precompute joins in advance, so that the expensive join operation is performed once and need not be repeated over and over again (which would be equivalent to what classical ILP systems do), but without actually storing the joint tables (which would increase its memory requirements). The second technique is a selective sampling method that can be used with rule learners that take the "covering" approach, learning one rule at a time and removing positive instances once they have been covered. The CrossMine algorithm is roughly an extension of Quinlan's FOIL system with these improvements added. Experimental results were presented, showing that CrossMine clearly outperforms other ILP systems with respect to efficiency, without suffering from an accuracy loss.

The next speaker was Adam Woznica, who presented *Kernel-based Distances for Relational Learning*, a paper by Adam Woznica, Alexandros Kalousis, and Melanie Hilario. With the surge of interest in kernel-based methods the last few years, several authors have recently researched how kernels can be defined for structural or relational data. This work continues in that direction: the authors define a family of kernel functions over relational schemata and use these kernel functions in an instance-based learner, thus obtaining a relational instance-based learner that performs well compared to previous approaches.

Finally, the paper *Dynamic Feature Generation for Relational Learning*, by Alexandrin Popescul and Lyle Ungar, was presented by the first author. The basic idea exploited in their paper is the following. The authors distinguish static feature generation, where features are generated from a predefined set in some predefined order, and dynamic feature generation, where features are generated according to information gathered during the search process. The authors' approach to dynamic feature generation is to predefine different sets of features (or "feature streams"), and let the learner generate new features from these sets taking into account which set has up till now yielded the most useful features. This approach was tested in an experiment on a citation database, using two streams of features, each of them using a different type of aggregate operator, and a significant efficiency improvement was found.

3. CONCLUSION

This workshop continued ongoing efforts in bringing together an international community that historically has been split across different conferences and workshops, and has thus reached one of its central goals: to further research on multi-relational and structural problems irrespective of origin and community. We certainly hope that the momentum gained by this third workshop will continue to foster close cooperation between all researchers interested in this topic from different perspectives.

This summary of course cannot do justice to the interesting contributions presented at the workshop. The reader is therefore encouraged to consult the workshop web site at `http://www-ai.ijs.si/SasoDzeroski/MRDM2004/`, where electronic copies of all papers can be found, as well as the presentations of the invited speakers.

Acknowledgements

Hendrik Blockeel is a post-doctoral fellow of the Fund for Scientific Research of Flanders, Belgium (FWO-Vlaanderen).

4. REFERENCES

[1] S. Džeroski, L. De Raedt, and S. Wrobel, editors. (2003) *Proceedings of the Second International Workshop on Multi-Relational Data Mining*. KDD-2003: Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC. `http://www-ai.ijs.si/SasoDzeroski/MRDM2003/`

[2] S. Džeroski, L. De Raedt, and S. Wrobel, editors. (2002) *Proceedings of the First International Workshop on Multi-Relational Data Mining*. KDD-2002: Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada. `http://www-ai.ijs.si/SasoDzeroski/MRDM2002/`

[3] S. Džeroski and N. Lavrač, editors. (2001) *Relational Data Mining*. Springer, Berlin.