Multirelational Data Mining 2003: Workshop Report

Saso Dzeroski
Dept. of Intelligent Systems
Jozef Stefan Institute
Ljubljana, Slovenia
Saso.dzeroski@ijs.si

Luc De Raedt
Machine Learning Lab
Albert-Ludwigs-Univ. Freiburg
Freiburg, Germany
deraedt@informatik.uni-freiburg.de

Stefan Wrobel
Fraunhofer AIS and Univ. of Bonn
Schloss Birlinghoven
Sankt Augustin, Germany
wrobel@ais.fraunhofer.de

ABSTRACT

In this report, we briefly review the second International Workshop on Multi-Relational Data Mining (MRDM-03), which was organized by the authors and held in Washington, D.C. on August 27th, 2003 as part of the workshop program of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03). 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.

Keywords

Multi-relational learning and data mining

1. INTRODUCTION

Multi-Relational Data Mining (MRDM) is the multi- disciplinary field dealing with knowledge discovery from relational databases consisting of multiple tables. 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. The field 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 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 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).

2. SUMMARY OF CONTRIBUTIONS

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. It was the second of its kind, following the first workshop on Multi-Relational Data Mining, held at SIGKDD 2002. In total, around 50 people attended the workshop which consisted of one invited presentation and 10 contributed papers which are summarized in the following.

In his invited talk, Raghu Ramakrishan set the scene for the subsequent presentations, identifying opportunities and challenges for the field from the perspective of databases. In the remaining contributed papers, it became clear that compared to the early days of multi-relational learning and learning in logic, and also in comparison to past year's workshop, the field has again broadened in scope and has already started addressing some of the challenges presented in the invited talk. The contributed papers can be grouped along four topics: the use of probabilistic models, scalability issues, propositionalization, and theory of multi-relational data mining.

2.1 Probabilistic Models

Even though the characteristic capability of multi-relational methods is being able to process multiple interrelated tables (first-order representation), it has become quite obvious in recent years how powerful these methods can become if probabilistic modelling capabilities are added. This line of research includes relational variants of different kinds of graphical models, as exemplified by several papers in the workshop.

A first contribution, *Prolog for First-Order Bayesian Networks: A Meta-interpreter Approach* (H. Blockeel), was concerned with multi relational extensions of Bayesian Belief networks. the paper demonstrates that using a meta-interpreter approach in Prolog, it is possible to represent such networks within the standard language of logic programming, thus offering a particular is simple point of comparison for other approaches to relational Bayesian networks have been proposed previously.

In *A Structural GEM for Learning Logical Hidden Markov Models* (K. Kersting, T. Raiko, L. De Raedt) the authors are concerned with another kind of graphical model, namely logical hidden Markov models. These models are capable of more compactly representing structured phenomenon, but their structure has so far proven difficult to learn. The authors present a novel structure learning algorithm based on generalized expectation maximization combined with structure search, and show its effectiveness in learning logical HMMs.

A third paper, *Collective Classification with Relational Dependency Networks* (J. Neville, D. Jensen), dealt with the task of Collective classification, i.e., exploiting the dependencies in a network to improve classification. In this context, the authors present relational dependency networks, a kind of undirected graphical model that extends prior work on simple dependency networks to make it applicable to the relational task of collective classification. In empirical experiments, this model indeed outperformed prior approaches such as relational probability trees.

Finally, even though it does not directly deal with probabilistic models, the paper *A Simple Relational Classifier* (S.A. Macskassy, F. Provost) provides an interesting counterpoint to the three other papers, and that it presents a very simple classifier, the Relational Neighbor (RN) classifier, that predicts only based on class labels of related neighbors, using no further learning. Interestingly, when compared to more complex approaches such as probabilistic relational models or relational probability trees, the RN approach perform surprisingly well and that suggests itself as a baseline method for judging more complex relational learners.

2.2 Scalability

Scalability to larger and more complex data sets has always been an issue in multi-relational learning, and is becoming even more important for very large challenge problems such as link discovery or discovery in biological domains. Approaches to scalability try to improve the basic algorithms, reduce the data in amount or complexity, or resort to propositionalization in order to use fast propositional learners.

In the paper *Efficient Multi-relational Classification by Tuple ID Propagation* (X. Yin, J. Han, J. Yang) the authors propose a very basic optimization technique for much relational learners. Given that multi-relational learning involves multiple joins between several tables, the cost of join operations may become a dominating element in runtime complexity. The authors address this point with a particular method of propagating tuple IDs to dependent relations, thus eliminating the need to actually perform joins during the learning run in certain cases, resulting and in very significant speed up.

The paper *Scaling Up ILP to Large Examples: Results on Link Discovery for Counter-terrorism* (L.R. Tang, R.J. Mooney, P. if Melville) addresses the issue of scaling to very large and complex background knowledge as it naturally arises in the task of link discovery. The paper shows that existing ILP (inductive logic programming) systems, in particular based on bottom-up search, are not well equipped to handling such tasks, and presents a novel algorithm that combines both top-down and bottom-up elements. On a standard benchmark problem, the new approach show significant gains in efficiency.

In *Towards feature selection for disk-based multirelational learners: a case study with a boosting algorithm* (S. Hoche, S. Wrobel), the authors examine an approach to feature selection in multi-relational learning that exploits the dynamics of a boosted learning algorithm. The approach monitors the development of boosting's classification margin, and brings in new features/relations whenever margin growth slows down. The paper shows that significant efficiency can be gained if introduction of features takes into account which relations have already been used, compared to strategy that simply orders features according to heuristic value.

2.3 Propositionalization

With respect to propositionalization, in *Structural logistic Regression for Link Analysis* (A. Popescul, L.H. Ungar), the authors show that logistic regression can successfully be upgraded to the relational setting by coupling it with relational feature generation and propositionalization. In their approach, propositionalization is not done a priori, but constitutes an integrated part of the algorithm's search. The paper shows that the resulting approach is suitable for the task of link prediction, in particular for predicting citations as found in CiteSeer.

In contrast, in *Constraint-Based Relational Subgroup Discovery* (F. Zelesny, N. Lavrac, S. Dzeroski), a preprocessing approach to propositionalization s taken. To make this approach feasible, it is based on a particular individual-centered representation which allows very detailed control over the propositionalization that is performed. The authors show that by incorporating additional constraints both into the generation of features and into the rule learning algorithm, the task of relational subgroup discovery can successfully be addressed.

2.4 Theory

Last, but certainly not least, the workshop has featured a theoretical contribution adding to the understanding of hypothesis and pattern spaces in much relational learning. In *Towards a Formal Framework for Mining General Patterns from Ordered Data* (G. Casas-Garriga) an analysis of mining order and collections of data such as sequential databases or time-series data is presented. The analysis is based on the idea of using closures of Galois connections taken from formal concept analysis. The paper contributes a new approach and shows that it is capable of exactly deriving the closed sequential patterns found by recent algorithm.

3. CONCLUSION

This workshop brought 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 second workshop will continue to foster close cooperation between all researchers interested in this topic from different perspectives.

This summary certainly cannot do justice to the interesting contributions that were presented at the workshop. The reader is therefore encouraged to consult the workshop web site at http://www-ai.ijs.si/SasoDzeroski/MRDM2003/, where electronic copies of all papers can be found.

.

4. ACKNOWLEDGMENTS

This paper was partially supported by DFG grants of the third author (WR 40/1-3, \Active Learning", and WR 40/2-1, “Hybrid Methods for heterogeneous information spaces”).

About the authors:

Saso Dzeroski 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/coeditor 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 ith 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 postdoctoral 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 chemoinformatics.

Stefan Wrobel is a professor of computer science at Univ. of Bonn and institute director at Fraunhofer AIS in Sankt Augustin/Bonn, Germany. Before moving to Bonn in 2002, he had been professor of computer science at Univ. of Magdeburg. He has been active in machine learning and data mining research since 1986. He has (co-)organized several conferences and workshops (e.g. ILP-94, ECML-95, LWA-99, MRDM-02). He is an action editor of the Journal of Machine Learning Research and an elected founding member of the International Machine Learning Society (IMLS). He is a member of the management board of KDnet, the European network of excellence on Knowledge Discovery, and has participated in several other European projects on machine learning and data mining. Among his research interests are scalability and local discovery, especially in multirelational learning.