Multi-Relational Data Mining 2005: Workshop Report

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

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

ABSTRACT

In this report we briefly review the 4th Workshop on Multi-Relational Data Mining (MRDM-2005), which was organized by the authors and held in Chicago, IL, on August 21, as part of the workshop program of the Eleventh 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 interested in methods and applications of finding patterns in expressive languages from multi-relational, complex, and/or structured data.

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 discussing practical applications of such techniques. MRDM-2005 was the fourth 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 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 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 fourth of its kind. It follows the success of the previous workshops on Multi-Relational Data Mining, held at SIGKDD 2002, 2003 and 2004, reports on which appear in SIGKDD Explorations [Vols 4(2), 5(2), 6(2)]. Further information on these workshops can be found at `http://www-ai.ijs.si/SasoDzeroski/MRDM200x/` (where x=2,3,4,5). Based on MRDM-02, a special issue of SIGKDD Explorations [Vol 5(1)] was co-edited by Sašo Džeroski and Luc De Raedt.

This year, some 40 people attended the workshop, which consisted of one invited presentation, seven full-length presentations and four short presentations. We briefly summarize them here.

The program started with a talk on "Gene classification: Issues and challenges for relational learning", authored by Claudia Perlich and Srujana Merugu, and presented by the latter. The authors investigate to what extent the structural and statistical properties of the domain of gene classification clash with the assumptions typically made by relational learning approaches. They propose a few potential solutions suggesting feature construction techniques that render the relational learner more suitable for a specific domain.

Indrajit Bhattacharya presented "Relational Clustering for Multi-type Entity Resolution", work in collaboration with Lise Getoor. Entity resolution refers to the problem of deciding whether two different objects actually refer to the same entity; for instance, a single author's name might occur in a bibliographic database in a number of variants. Besides properties of the object itself, properties of its environment (other objects related to it) are important to perform entity

resolution, which makes this a typical relational learning problem. Bhattacharya and Getoor study the problem of entity resolution in a context where objects are of different types. They show how the problem can be approached by posing it as a relational clustering problem.

The next talk was by Matthew Rattigan, who presented "The case for anomalous link detection", work done together with David Jensen. In this presentation, important challenges inherent to link prediction were discussed, which relate mainly to the enormous class skew (the actual number of links is typically much smaller than the number of nodes squared). The alternative and complementary task of Anomalous Link Discovery (ALD) was presented. ALD considers the task of detecting, among existing links, which ones are unexpected. The authors show that ALD is much simpler than link prediction and also has numerous interesting applications, and argue that more research into ALD is desirable.

After the morning coffee break, Jennifer Neville talked on "Leveraging Relational Autocorrelation with Latent Group Models", again work with David Jensen. Autocorrelation is a statistical dependency between the values of the same variable on related entities. The phenomenon is ubiquitous in relational datasets, and previous work by these authors has shown that it can both confuse learners that ignore this phenomenon, and boost the performance of techniques that explicitly exploit it (e.g., collective classification methods). In this paper, the authors argue that a common cause of autocorrelation is the existence of underlying (latent) groups among the objects, and that methods that explicitly look for such latent groups might exhibit better performance, allow more efficient inference, and yield more insight in the dataset. The authors propose a latent group model for relational data and experimentally show that it indeed outperforms models that ignore latent group structure.

Christine Körner presented the paper "Bias-free hypothesis evaluation", by herself and Stefan Wrobel. The starting point of this work is the earlier work by Jensen and colleagues on the effects of relations among objects on relational learning, and more specifically, on biases introduced in test-set based evaluation that are due to test set examples being linked to training set examples. Körner and Wrobel discuss that the solutions proposed previously discuss one specific case of a more general problem. They propose an algorithm, *generalized subgraph sampling*, that avoids bias in the test set also for the more general case.

The next presentation was given by Hongyu Guo, together with Herna Viktor author of "Mining Relational Databases with Multi-view Learning". They present an approach to mining relational databases that they call multi-view learning, and the essence of which is the following. Data in a relational database generally cannot be reduced to a single table without loss of information. However, one can define multiple views that gather data from different relations, where each view keeps different pieces of information. As the result of each view is a single table, each of them can be mined using standard non-relational learners, after which the results of these learners can then be combined in a meta-learning step. The authors present promising experimental results for this method.

The last talk of the second session was given by Irene Ong, presenting work in collaboration with David Page, Inês Dutra and Vítor Santos Costa: "HyperPaths: Extending Path-Finding to Moded Languages". In the early days of Inductive Logic Programming, the technique of Pathfinding was proposed by Richards and Mooney to efficiently search the enormous hypothesis spaces encountered in ILP. Ong et al. now pick up this thread to improve present-day ILP systems. Several of those, such as Aleph and Progol, look for hypotheses by first constructing a so-called bottom clause, which expresses everything that is known about a single example, and then look for subsets of this bottom clause, which express more general rules that still cover this example. Ong et al. extend pathfinding to make use of mode declarations, and apply this to find paths in the bottom clause that lead to a good solution. They show experimentally that this technique leads to the discovery of interesting clauses that could not easily be found with Aleph's standard search procedure.

After a quick lunch on a sunny terrace near Chicago's lakefront, David Page opened the afternoon session with a talk on "Learning Bayesian Networks of Rules with SAYU", by Jesse Davis, Elisabeth Burnside, David Page, Inês Dutra and Vítor Santos Costa. Whereas standard ILP methods typically create sets of if-then rules where each rule in itself is assumed to be correct, Davis et al. explore methods to combine the predictions of such rules in a less straightforward fashion, for instance using a bayesian classifier. Specific for the work presented here is the idea that rules should not be learned using a heuristic that is independent of the way the rules are going to be used. Standard heuristics for learning rules assume that each rule needs to achieve maximal correctness by itself, but if a rule's prediction will be used as part of a bayesian net, then these heuristics may not be optimal. Davis et al. propose a method where rules are scored by evaluating how well they fit into a bayesian network that is being created, a method they call SAYU: Score As You Use. Experiments show that the method tends to yield theories that are simpler and have higher predictive performance than when the original heuristics are used, and moreover fewer rules need to be evaluated to achieve this goal.

The next presentation was by Hongyan Liu, Xiaoxin Yin and Jiawei Han: "An efficient Multi-relational Naive Bayesian Classifier Based on Semantic Relationship Graph". This research is in the line of other approaches such as Flach and Lachiche's 1BC and Ceci et al.'s Mr-SBC, who all studied how the Naive Bayes learning method can be upgraded to the first-order or multi-relational setting. Liu et al. propose a novel method that is based on the concept of a semantic relationship graph, which indicates which attributes naturally link different tables together (similar to the concept of foreign key relationships), and helps to avoid meaningless joins. They further argue that just joining multiple tables into one table may cause replication of attribute values in multiple tuples of the result, thus unduly increasing the weight of some attributes, which distorts the classifier; their method avoids this. Experiments confirm their method has high efficiency and good accuracy.

Gerson Zaverucha presented "Further Results of Probabilistic First-order Revision of Theories from Examples", by Aline Paes, Kate Revoredo, Gerson Zaverucha and Vítor Santos Costa. Theory revision has been an important topic in inductive logic programming for many years, and with the advent of formalisms combining probabilistic reasoning with first order logic, it is not a big surprise that researchers take up the challenge of studying theory revision in these

settings. Paes et al. study theory revision in the context of Kersting and De Raedt's Bayesian Logic Programs. They present the PFORTE system and discuss experimental results with it, confirming the expected advantages of revising an approximately correct probabilistic theory over learning a theory from scratch.

Nikhil Ketkar presented the paper "Qualitative Comparison of Graph-based and Logic-based Multi-relational Data Mining: A Case Study", work in collaboration with Lawrence Holder and Diane Cook. Logic-based and graph-based methods are two different kinds of approaches towards learning from structured data, each with their own strengths and weaknesses. While there has been speculation regarding these strengths and weaknesses, this is probably the first time that both approaches are compared in a single case study. A experimental comparison between the graph-based learner Subdue and the ILP system CProgol indicates that Subdue is better at discovering structurally large concepts, whereas CProgol has advantages for learning semantically complicated concepts and can make better use of background knowledge.

The workshop ended with an invited talk by Thomas Gärtner, from the Autonomous Intelligent Systems lab of the Fraunhofer Gesellschaft für Informatik, Germany. Much of Thomas Gärtner's work focuses on kernel methods for structured data. In this presentation, he gave an overview of his work on kernel methods for graphs, mostly on an intuitive introductory level, but also regularly zooming in on technical details. Besides explaining how one can devise kernel functions that express the similarity between graph structures in some meaningful way, Gärtner also discussed the efficiency of these methods, showing that some of these scale sufficiently well to be applicable on massive graphs. The latter is an important requirement for applications such as Web mining.

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 fourth workshop will continue to foster close co-operation between all researchers interested in this topic from different perspectives.

4. ACKNOWLEDGEMENTS

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