A Report on the Summer School on Relational Data Mining

17-18 August 2002, Helsinki, Finland

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

Bernard Ženko
Jozef Stefan Institute
Jamova 39
SI-1000 Ljubljana, Slovenia
bernard.zenko@ijs.si

Relational Data Mining (RDM) is the multi-disciplinary field dealing with knowledge discovery from relational databases consisting of multiple tables. To emphasize the contrast to typical data mining approaches that look for patterns in a single relation of a database, the name Multi-Relational Data Mining (MRDM) is often used as well. Mining data which consists of complex/structured objects also falls within the scope of this field: 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 (ILP), KDD, data mining, machine learning and relational databases; producing new techniques for mining multi-relational data; and practical applications of such techniques.

Present RDM 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. RDM 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. RDM 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 (e.g., information extraction from Web sources).

The Summer School on Relational Data Mining provided a comprehensive introduction to the techniques and applications of relational data mining by leading experts in the field. The lectures given at the school are summarized below. The slides of the lectures were published as handouts and are also available for download at the Web page of the school http://www-ai.ijs.si/SasoDzeroski/RDMSchool

- The introductory lecture,"An introduction to relational data mining" by Sašo Džeroski, first summarized the standard data mining tasks (classification, regression, clustering, association discovery) and approaches (trees, rules, nearest-neighbor) and illustrated them for the propositional case. After demonstrating the need to mine (multi)-relational data, the topic of RDM was introduced and a brief overview of RDM approaches was given.

- The second lecture,"An introduction to inductive logic programming", given by Peter Flach on behalf of Nada Lavrač, introduced the field of ILP, which is concerned with learning logic programs from examples and background knowledge. Fundamental topics such as searching the space of program clauses and the generality lattice induced by theta-subsumption were covered, as well as generic ILP approaches based on the search of refinement graphs and least general generalization.

- The lecture "Propositionalization as a way of understanding RDM and ILP", also presented by Peter Flach, covered another generic approach to ILP, namely the approach of transforming a relational learning problem to a propositional one. The talk clarified the relationship between relational and propositional learning and the role of representation in RDM. It also presented an approach to propositionalization for individual-centered strongly typed representations, where feature construction is guided by types.

- The three main messages of the lecture "A methodology of ILP" by Luc De Raedt were that: (1) ILP applies essentially to any machine learning / data mining task, not just concept learning, (2) There is a recipe for deriving new ILP algorithms from propositional ones, and (3) ILP as an expressive framework has many special cases, which are often studied separately.

- The lecture "Logical trees for classification, regression and clustering" by Hendrik Blockeel presented an approach to using decision trees for relational problems. The approach extends well known techniques for induction of decision trees in two ways: (1) Predictive clustering generalizes over several induction tasks (classification regression and clustering), while (2) First order logical decision trees make trees usable in the

context of ILP. The resulting technique therefore combines the high expressiveness of first order logical decision trees with the efficiency and accuracy of decision trees.

- The lecture "Relational subgroup discovery", given by Stefan Wrobel, dealt with the problem of finding interesting patterns which are only valid in selected regions of the instance space. These local patterns are often impossible to find if we try to model the entire instance space. The talk first introduced several medical and business problems which can be successfully solved by relational subgroup discovery, and then presented the algorithm MIDOS which can be used for finding relational descriptions of subgroups.

- The next lecture by Stefan Wrobel, titled "Relational distance-based methods", presented several relational instance based learning systems and explained the distance functions used in these systems. The talk also presented several successful applications of these techniques in the fields of chemistry and biology.

- The lecture "Kernel-based learning from structured data", given by Thomas Gaertner, first gave a clear explanation of the basics of support vector machines. The problem of selecting an appropriate kernel function for a given domain was pointed out, emphasizing that the kernel is a part of the background knowledge about a domain. The talk then discussed how to construct kernel functions for structured data and concluded with some example applications of this technique.

- The lecture "Learning statistical models from relational data" given by Lise Getoor presented an extension of statistical models for relational problems. It first covered the topic of Bayesian networks, which are the foundation of the Probabilistic Relational Models (PRMs). PRMs, ways to learn their structure and parameters from data, and their applications to practical problems were then discussed.

- "Bayesian logic programs" (BLPs), discussed in the lecture by Luc De Raedt (presenting joint work with Kristian Kersting), are the first order equivalent of Bayesian networks. In addition, they generalize pure Prolog, dynamic Bayesian netorks, dynamic Bayesian multinets, hidden Markov models etc. The talk first focussed on the language of BLPs, then on leaning the parameters and structure of BLPs.

- In the lecture "Applications of ILP/RDM to bioinformatics", Ross King presented a number of successful applications of ILP/RDM to real problems in the field of bioinformatics. The applications presented included structure activity relationship modeling, functional genomics and "The Robot Scientist Project.

- The lecture "RDM Applications; An overview", given by Sašo Džeroski, presented a variety of ILP/RDM applications with a more detailed description of some selected applications. Besides the applications in the area of bioinformatics, which were the topic of the previous talk, the lecture presented applications from areas of medicine, environmental science, traffic engineering, mechanical engineering, text/web mining, natural language processing, business data analysis, music, software engineering and adaptive systems management.

- The last lecture on "Inductive databases" was given by Luc De Raedt. Inductive databases store both data and patterns valid in the data. In inductive databases, data mining is viewed as a querying process. The first part of the talk presented several examples of (preliminary) inductive databases while the second part described a logic oriented view on the principles of inductive databases.

The Summer School concluded with a panel discussion, where the lecturers and participants raised a number of interesting issues concerning the future of relational data mining. Interesting statements from the panel include: "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)"; "The different approaches to RDM should be integrated" (Hendrik Blockeel); "RDM should be more of an engineering discipline than art" (Thomas Gaertner); "The European and US scientific communities working on RDM topics should communicate and not develop in isolation" (Lise Getoor).

The Summer School was organized by the Jozef Stefan Institute,
Ljubljana, with the help and support of the University of Helsinki. It was financially supported by ILPnet2 (The Network of Excellence in Inductive Logic Programming). It was attended by 34 participants from 13 countries (including the USA and New Zealand).