Multi-Relational Data Mining: The Current Frontiers

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

Luc De Raedt
Institut für Informatik, Albert-Ludwigs-University,
Georges-Koehler-Allee, Building 079, D-79110
Freiburg, Germany
deraedt@informatik.uni-freiburg.de

Data mining algorithms look for patterns in data. Most existing data mining approaches are propositional and look for patterns in a single data table. Most real-world databases, however, store information in multiple tables.

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) (Džeroski et al. 2002). We will adopt this term in the present special issue.

When we are looking for patterns in multi-relational data, it is natural that the patterns involve multiple relations. They are typically stated in a more expressive language than patterns defined on a single data table. The major types of multi-relational patterns extend the types of propositional patterns considered in single table data mining. We can thus have multi-relational classification rules, multi-relational regression trees, and multi-relational association rules, among others.

Just as many data mining algorithms come from the field of machine learning, many MRDM algorithms come form the field of inductive logic programming (ILP, Muggleton 1992; Lavrač and Džeroski 1994). Situated at the intersection of machine learning and logic programming, ILP has been concerned with finding patterns expressed as logic programs. Initially, ILP focussed on automated program synthesis from examples, formulated as a binary classification task. In recent years, however, the scope of ILP has broadened to cover the whole spectrum of data mining tasks (classification, regression, clustering, association analysis). The most common types of patterns have been extended to their multi-relational versions and so have the major data mining algorithms (decision tree induction, distance-based clustering and prediction, etc.).

There is also a growing interest in the development of data mining algorithms for various types of structured data. These include, for example, graph-based data mining. There is also an increasing body of work on mining tree-structured and XML documents. Mining data which consists of complex/structured objects also falls within the scope of MRDM, as the normalized representation of such objects in a relational database requires multiple tables.

The rise of several KDD application areas that are intrinsically relational has provided and continues to provide a strong motivation for the development of MRDM approaches. These include in the first place bioinformatics and more broadly computational biology. The World Wide Web and data mining tasks related to it, such as information extraction follow closely. Related tasks, such as social network analysis also hold significant importance.

In this special issue, we first provide an introduction to multi-relational data mining (Džeroski 2003). An extensive overview of the field is given by Džeroski and Lavrač (2001). The overview in this issue focusses on the most important approaches, giving a brief introduction to inductive logic programming, then covering several multi-relational data mining tasks and techniques, such as multi-relational association rules and multi-relational decision trees.

Most of the full-length contributions in this special issue cover important recent advances at the frontiers of multi-relational data mining. These include probabilistic logic learning (De Raedt and Kersting 2003), kernel-based learning for structured data (Gärtner 2003), and graph-based data mining (Washio and Motoda 2003). Blockeel and Sebag (2003) survey scalability and efficiency issues in MRDM and approaches taken to resolving these. Finally, Page and Craven (2003) survey several applications of MRDM in the area of bioinformatics, the area where MRDM has had most successes so far and which still holds many challenges for MRDM.

In addition to the six full-length contributions, this special issue contains three shorter articles/position statements. While Washio and Motoda (2003) mainly survey methods for frequent subgraph discovery, Holder and Cook (2003) discuss current and future directions in graph-based relational learning, which also addresses other data mining tasks, such as classification. Getoor (2003) introduces and discusses several tasks of link mining, where links/relations among objects are of central importance. Finally, Domingos (2003) discusses several applications areas for MRDM, as well as challenges that MRDM must address to be successful in these.

A summer school on relational data mining, covering both basic and advanced topics in this area, was organized in Helsinki in August 2002, preceding ECML/PKDD 2002 (The 13th European Conference on Machine Learning and The 6th European Conference on Principles and Practice of Knowledge Discovery in Databases). The slides from this summer school are available online (Džeroski and Ženko 2002). A report on this event by Džeroski and Ženko (2003) is included at the end of this special issue.

1. REFERENCES

[1] H. Blockeel and M. Sebag. Scalability and Efficiency in Multi-Relational Data Mining. *This issue.*

[2] L. De Raedt and K. Kersting. Probabilistic Logic Learning. *This issue.*

[3] P. Domingos. Prospects and Challenges for Multi-Relational Data Mining. *This issue.*

[4] S. Džeroski. (2003) Multi-Relational Data Mining: An Introduction. *This issue.*

[5] 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.

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

[7] S. Džeroski and B. Ženko, editors. (2002) Lecture Notes of the Summer School on Relational Data Mining, 17-18 August 2002, Helsinki, Finland. `http://www-ai.ijs.si/SasoDzeroski/RDMSchool/`

[8] S. Džeroski and B. Ženko. (2003) A Report on the Summer School on Relational Data Mining, 17-18 August 2002, Helsinki, Finland. *This issue.*

[9] T. Gärtner. (2003) Kernel-based Learning in Multi-Relational Data Mining. *This issue.*

[10] L. Getoor. (2003) Link Mining. *This issue.*

[11] L. Holder and D. Cook. (2003) Graph-based Relational Learning: Current and Future Directions. *This issue.*

[12] N. Lavrač and S. Džeroski. (1994) *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, Chichester, 1994. Freely available at `http://www-ai.ijs.si/SasoDzeroski/ILPBook/`.

[13] S. Muggleton, editor. (1992). *Inductive Logic Programming*. Academic Press, London.

[14] D. Page and M. Craven. (2003) Biological Applications of Multi-Relational Data Mining. *This issue.*

[15] T. Washio and H. Motoda. (2003) State of the Art of Graph-based Data Mining. *This issue.*