Fundamenta Informaticae 89 (2008) i-iii
IOS Press

Progress on Multi-Relational Data Mining

Preface

Multi-Relational Data Mining (MRDM) is an emerging research area dedicated to the development and application of novel computational techniques for knowledge discovery from relational databases, where the description of units of analysis is spread over 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.

MRDM has a precursor going back over a decade in the field of Inductive Logic Programming (ILP). Situated at the intersection of machine learning and logic programming, ILP has been concerned with finding patterns expressed as logic programs. Although ILP was initially focused on automated program synthesis from examples, the scope of ILP has recently broadened to cover the whole spectrum of data mining tasks (classification, regression, association analysis, clustering). The most common types of patterns and data mining algorithms have been extended to the multi-relational case and MRDM studies now encompass relational association rule discovery, relational classification rules, relational decision and regression trees and probabilistic relational models among others. At same time, MRDM methods have been successfully applied across many application areas, ranging from the analysis of business data, through bioinformatics and pharmacology to Web mining and Spatial Data mining.

Although MRDM has reached a relative maturity over the last years, we are far from a deep understanding of the issues in MRDM. There is still a number of interesting and open questions. For instance, one of the central research topics of MRDM is concerned with combining expressive knowledge representation formalisms such as relational and first-order logic, with principled probabilistic and statistical approaches to inference and learning. This combination is needed in order to face the challenge of real-world learning and data mining problems, in which the data are complex and heterogeneous and we are interested in finding useful predictive and/or descriptive patterns. Further research issues are related to the limited scalability of multi-relational algorithms, the management of data with a spatio-temporal semantics, the high variability in accuracy of existing algorithms, and the identification of killer applications.

For this special issue, we have selected seven papers which constitute an updated and extended version of the papers presented at the Workshop on Multi-Relational Data Mining, Warsaw (Poland), 17th September 2007, chaired by both the guest editors of this special issue and Dr. Michelangelo Ceci (University of Bari, Italy). The workshop was a satellite event of the 18th European Conference on Machine Learning (ECML 2007) and the 11th European Conference on Principles and Practice of

Knowledge Discovery in Databases (PKDD 2007). All papers selected for this special issue have been peer-reviewed and appropriately revised as required by this journal.

To facilitate orientation for the reader of this special issue, we summarize the main ideas behind the individual papers in the following:

The first paper, ,,An Experimental Comparison of Different Inclusion Relations in Frequent Tree Mining" by J. De Knijf and A. Feelders compares a variety of algorithms to derive all frequent sub-trees from a database of labeled ordered rooted trees. The fundamental difference between these algorithms lies in the different notions of when a tree matches another tree and how attributes are handled. Different algorithms are compared with respect to the usefulness of the derived patterns and, in particular, the performance of classifiers which use the derived patterns as features. Comparison is based on time and memory performance. Experiments on both artificial and real databases confirm that a significant improvement in both predictive performance and computational efficiency can be gained by choosing the right tree mining approach.

In ,,Multi-Dimensional Relational Sequence Mining", F. Esposito, N. Di Mauro, T.M.A. Basile and S. Ferilli present an ILP algorithm to mine complex patterns, expressed in a first-order language, in which events may occur along different dimensions (e.g. spatio-temporal sequences). Multidimensional patterns are defined as a set of atomic first-order formulae, in which events are explicitly represented by a variable and the relations between events are represented by a set of dimensional predicates. Experiments are performed on artificial and real multi-dimensional sequences.

The efficiency of ILP systems when applied to complex problems is investigated in the paper by N.A. Fonseca, R. Camacho, R. Rocha and V.S. Costa, entitled ,,Compile the hypothesis space: do it once, use it often". The authors propose a novel technique which avoids deducing each example to evaluate every constructed clause. The technique is based on the Mode Directed Inverse Entailment approach to ILP, where a bottom clause is generated for each example and the generated clauses are subsets of the literals of this bottom clause. Clauses generated from all bottom clauses are stored in a prefix-tree, together with some extra information. This information is shown to be sufficient to estimate the number of examples which can be deduced from a clause. In addition, the authors present an extension of this algorithm, where each prefix-tree is computed only once (compiled) per example. Both algorithms are evaluated on real applications and considerable speedups are observed.

The paper ,,Learning from Skewed Class Multi-relational Databases" by H. Guo and H. L. Viktor presents a strategy to deal with imbalanced multirelational data. The method learns from multiple views (feature sets) of relational data and constructs view learners with different awareness of the imbalanced problem. These different observations, collected by the multiple view learners, are combined in order to yield a model which has better knowledge of both the majority and minority classes in a relational database. Experiments on benchmark datasets show that the proposed method achieves promising results when compared with other popular relational data mining algorithms.

The paper ,,A Restarted Strategy for Efficient Subsumption Testing" by O. Kuzelka and F. Zelezny focuses on the problem of verifying the subsumption relation between a relational pattern and an example. The authors design a randomized restarted modification, called ReSumEr, of the baseline algorithm which is complete and avoids heavy tails of the baseline algorithm's runtime distributions. The algorithm is further augmented into ReSumEr2 by allowing transfer of information among individual restarts. ReSumEr2 has been tested against the state-of-the-art subsumption algorithm Django on both generated graph data and the predictive toxicology challenge (PTC) data set.

The sixth paper is entitled ,,Relational Transformation-based Tagging for Activity Recognition" by N. Landwehr, B. Gutmann, I. Thon, L. De Raedt and M. Philipose. It presents a relational transformation-based tagging system based on ILP principles, which is able to cope with expressive relational representations as well as a background theory. The approach is experimentally evaluated and compared to Hidden Markov Models on two activity recognition tasks, as well as an information extraction task.

The last paper, ,,Learning Ground CP-Logic Theories by Leveraging Bayesian Network Learning Techniques" by W. Meert, J. Struyf and H. Blockeel is related to the Causal Probabilistic Logic (CP-logic), that is a probabilistic modeling language especially designed to express causal relations. The authors propose an algorithm, called SEM-CP-logic, which applies Bayesian network (BN) learning techniques to learn a CP-theory in the form of an equivalent BN. The work also provides a theoretical and experimental comparison between CP-theory and BN learning. The comparison shows that the most simple CP-theories can be represented with BNs, consisting of noisy-OR nodes, while more complex theories require amost fully connected networks (unless additional unobserved nodes are introduced in the network). SEM-CP-logic is used in a medical application in the context of HIV research.

The seven papers in this special issue constitute only a small sample of recent research related to multi-relational data mining. However, they give the reader a sense of some of the most challenging problems which arise in multi-relational data mining. We hope that this issue will draw added attention to this field and stimulate further research.

Finally, we want to thank the diligent reviewers for their critical comments on these manuscripts and the authors for their timely and careful revisions. Most of all, we hope you enjoy reading these articles and find the content interesting and useful in your work.

Special Issue Editors

Donato Malerba
Annalisa Appice
Dipartimento di Informatica,
Università degli Studi di Bari, Italy
{malerba, appice}@di.uniba.it