Mach Learn (2006) 62: 3–5
DOI 10.1007/s10994-006-5856-7

EDITORIAL

Introduction to the special issue on multi-relational data mining and statistical relational learning

Hendrik Blockeel · David Jensen · Stefan Kramer

Within the field of machine learning, it has long been recognized that the difficulty of learning is related, among other things, to the representational complexity of the knowledge that is input to the learner. Many machine learning approaches learn from examples that have a relatively simple, "standard", representation, and that are drawn independently from an identical distribution ("i.i.d."). In many application domains for machine learning, however, individuals may have a richer internal structure than the standard representation allows us to represent, and there may be relationships between them that invalidate the i.i.d.-assumption. As a result, standard learning techniques may not work very well in these domains.

More specifically, the majority of learning approaches have considered cases where each individual data element is described using a fixed set of attributes, each of which has a simple domain that contains atomic (non-decomposable) values, which may be continuous or discrete. But data, or knowledge in general, is not always available in this restricted format, nor can it always easily be converted into it. Individual data elements might be described in a way that is inherently more complex, for instance, as sets or graphs. Such structures can be summarized by listing a number of properties (the size of a set, number of nodes or edges in a graph, or in general certain distributional characteristic of these sets or graphs), but such a summary is always an abstraction of the actual information and as such gives rise to loss of information.

Note that this complex description need not describe the "internal structure" of an example. There are many cases where individuals have a simple internal structure, but are related to each other or to other external objects. If we want to take these relationships into account when learning, the description of such an individual includes its local environment and therefore becomes relational.

Ignoring this relational structure not only causes potentially useful information to be ignored; it may actually be harmful to the learner. For instance, it has been shown that the violation of the i.i.d.-assumption causes the standard heuristics used by learners to be misguided.

H. Blockeel · D. Jensen · S. Kramer
Guest editors

Relational learning explicitly aims at processing data in the form of sets, graphs, or similarly complex structures, where the standard representation and/or the i.i.d.-assumption are not used. It has its roots in several research communities.

From the early days of machine learning, researchers have looked into learning in "knowledge-rich domains", learning from structured data, or learning from first order logic descriptions of data. From the nineties onwards, much of this research has been performed under the umbrella of "inductive logic programming", which has resulted in a strong focus on logic formalisms for relational learning.

While the above developments happened in the area of symbolic learning, also on the side of probabilistic and statistical approaches interest in learning from structural and relational data has been growing. A well-known development in this area is, for instance, the probabilistic relational models developed by Daphne Koller and colleagues.

Finally, there are a number of more specialized areas, such as web mining, social networks analysis, and temporal and spatial analysis, which require relational learning approaches, and provide both motivation for and specific contributions to the area of relational learning.

Seeing a proliferation of techniques and approaches originating in different fields, each with their own strengths and weaknesses, it has increasingly been recognized that an integration of all these different approaches should be strived for. Researchers have been trying to combine the expressiveness of first order logic with the robustness of probabilistic reasoning, the orientation towards structure that graph based methods exhibit, and the efficiency of relational databases.

Many of these developments have been visible at a variety of conferences, including the major machine learning, data mining, and artificial intelligence conferences. The workshop series on Statistical Relational Learning (SRL) that started in 2000 and the workshop series on "Multi-Relational Data Mining" (MRDM) that started in 2002 can be considered dedicated to these developments. One could argue that the SRL series takes a more statistical viewpoint, whereas the MRDM series stands closer to the relational database viewpoint, but there is a large overlap in the coverage of these workshop series, in principle and in practice.

After several years of succesful SRL and MRDM workshops, many ideas presented there have matured, and the time seemed right for a special issue on the subject. This issue presents a selection of articles in the areas of statistical relational learning and multi-relational data mining. Several of them have grown out of earlier presentations at SRL or MRDM workshops. They illustrate the large variety of approaches that currently exist in relational learning, but also the extent to which this domain has matured.

Lise Getoor and John Grant introduce PRL, which stands for Probabilistic Relational Language. PRL can be seen as a recasting of probabilistic relational models (PRMs) into a logic programming framework. By doing this, the authors contribute to a better understanding of the relationships between the many different formalisms that currently exist for upgra ding Bayesian networks towards first order logic. More specifically, relationships between the closely related frameworks of Bayesian Logic Programs, Logical Bayesian Networks and PRMs are explicitly discussed and clarified.

Nada Lavrač and Filip Železny propose a new propositionalization approach to subgroup discovery in relational databases. The first stage of the approach enumerates non-redundant features, while the second uses a novel rule evaluation measure WRAcc and a weighted covering approach. One of the pruning criteria applied in the first stage is the non-decomposability of structural features.

Claudia Perlich and Foster Provost present a propositionalization approach that extends traditional aggregation operators like COUNT and MEAN with class-conditional Bayesian aggregates. It was developed for learning from networked domains, and for the aggregation of categorical attributes with high cardinality and in particular object identifiers.

Matthew Richardson and Pedro Domingos present the concept of Markov Logic Networks as a novel formalism that combines first order logic reasoning with probabilistic and uncertainty reasoning. They extend the existing line of research on probabilistic logic learning, which includes approaches such as Bayesian logic programs, stochastic logic programs, and independent choice logic, with a powerful yet intuitive new approach towards representing, exploiting, and learning logic formulae and the degree to which they may be violated.

Finally, Mohammed Zaki and Charu Aggarwal present a method for classification of tree structures, such as XML documents, based on their structural properties. Their XRules algorithm looks for tree structures that frequently occur embedded in the tree-structured data and that carry information about the class of the tree structure, and then constructs classification rules based on these substructures.

We hope that the forelying issue will not only contribute to a better understanding of the current state of the art in the areas of multi-relational data mining and statistical relational learning, but also increase the reader's interest in the many challenges that are still ahead of us.