IOS Press

Advances in Mining Graphs, Trees and Sequences

Preface

Ever since the early days of machine learning and data mining, it has been realized that the traditional attribute-value and item-set representations are limited for many practical applications in domains such as chemistry, biology, network analysis and text mining. This has triggered a lot of research on mining and learning within alternative and more expressive representation formalisms such as computational logic, relational algebra, graphs, trees and sequences. The state-of-the-art is that attribute-value and item-set representations lie at one end of the spectrum [1, 2], and multi-relational data mining and inductive logic programming at the other end [3, 4]. The middle is occupied by traditional data structures employed throughout the field of computer science. These include graphs, trees and sequences (or strings). The motivation for using such representations is that they are 1) more expressive (and therefore more widely applicable) than flat representations, and 2) potentially more efficient than multi-relational learning and mining techniques. At the same time, the data structures of graphs, trees and sequences are among the best understood and the most widely applied representations within computer science. Thus these representations offer ideal opportunities for developing interesting contributions in data mining and machine learning that are both theoretically well-founded and widely applicable.

Whereas there have been a large number of workshops and conferences devoted to multi-relational data mining and inductive logic programming as well as applications of intermediate representations in *e.g.* ontologies, bioinformatics, XML-data, text-mining, – to the best of our knowledge – the first scientific event specifically devoted to using intermediate representations in data mining was the *First International Workshop on Mining Graphs, Trees and Sequences (MGTS-2003)* [5]. This workshop provided a stimulating environment for exchanging information among researchers in this newly emerging sub-area of data mining. Its success motivated us to plan this special issue of *Fundamenta Informaticae* devoted to the state-of-the-art in mining graphs, trees and sequences. A considerable number of papers from all over the world has been submitted to the special issue, some of which were extensions of papers presented at the workshop. All submissions went through a strict reviewing process and only the best papers have been selected for inclusion in this special issue. They either provide an original survey of the field or else advance the state-of-the-art in this newly emerging research field. At this point, the editors also wish to point out that the views, opinions and motivations expressed concerning the applications and application domains in this special issue are those of the authors and should not be interpreted as those of the editors, neither expressed nor implied.

There are seven papers in this special issue. The first four papers are devoted to the discovery of substructures in sequences, trees and graphs. The fifth and sixth paper are concerned with the principles of modeling structured regularities that occur in structured data. The last paper gives an overview of tree mining algorithms and identifies theoretical foundations. We next discuss the papers in more detail.

1. The first paper by Sebastien Ferre and Ross D. King proposes an approach to model elementary data structures through the composition of logical components. Complete, non-redundant, flexible and efficient composition is achieved by a data-driven framework consisting of concept-based search, which focuses on the concepts to be discriminated, and dichotomic search, which explores the search space in both directions. The advantages of this approach are that it can cope with infinite chains in the search space and that is allows for a wide range of logics. The search for motifs in sequences is illustrated using the East-West train challenge and the problem of discriminating biological functions from Yeast protein secondary structures.

2. The second paper by Mohammed J. Zaki proposes an efficient algorithm, called SLEUTH, for mining frequent, unordered, embedded subtrees in a database of labeled trees. An equivalence class extension scheme, called Canonical Extension, is introduced in addition to the Prefix Extension for the complete generation of all candidate trees. The notion of scope-list joins is extended to compute the frequency of unordered trees. The performance is not only evaluated on several synthetic data set, but also on a real-world data set consisting of processed webserver access log files of one particular website. This data set will also be used in other papers to set all algorithms into a perspective. The evaluation in this paper shows that SLEUTH is efficient and its computation time is comparable to another algorithm, TreeMiner, which mines only ordered trees.

3. The third paper by Akihiro Inokuchi et al. proposes a general framework called Biased Apriori-based Graph Mining (B-AGM) to conduct a complete search for various classes of frequent subgraphs, *e.g.*, induced and connected subgraphs, unordered and ordered subtrees and paths embedded in a data set of labeled graphs. This work extends the AGM algorithm for inducing subgraphs, by introducing new bias constraints to limit the search space efficiently to an objective class of frequent subgraphs. The performance has been evaluated through real world graph and tree data sets of Predictive Toxicology Evaluation, anti-HIV activity and the web browsing access log data set. The evaluation shows the scalability and flexibility of this approach with respect to the amount of data and the computation time.

4. The fourth paper by Lawrence Holder et al. describes Graph-based Relational Learning as implemented in the Subdue system and its application to detecting security threats. It reviews some techniques of relational learning in graph-based data mining, and defines Graph-based Relational Learning focusing on identifying novel, but not necessarily maximally frequent, patterns in a graph representation of data. Then, several approaches encompassed by Subdue are explained. These include: substructure discovery, graph-based clustering, supervised learning and graph grammar learning. The application of Subdue to mining terrorist networks is demonstrated on the data of various threat and non-threat groups simulated in the EAGLE simulator of the U.S. Air Force program. The results indicate that Subdue is effective and efficient in learning patterns to distinguish threats from non-threats, especially when focusing on groups and communications between group members.

5. The fifth paper by Amaury Habrard et al. proposes a probabilistic approach that aims at a priori pruning noisy or irrelevant subtrees in a set of trees to improve probabilistic learning. Only a part of a tree, *i.e.*, a subtree, can be deleted rather than the whole tree itself. This method is based on a partitioning of the whole set of subtrees, using regular tree patterns or contexts, and on the evaluation of the relevance of the probability of a subtree to be in a given partition. Subtrees with a too small probability are deleted. This approach is evaluated in learning stochastic tree automata through several synthetic data set, a UCI bechmark data set and two real world data set converted into trees. These experiments show the efficiency and robustness of the approach: automata closer to the target concept can now be inferred in the presence of noisy data.
6. The sixth paper by Warodom Geamsakul et al. proposes a method called Decision Tree Graph-Based Induction (DT-GBI), which constructs a decision tree classifier for graph-structured data while simultaneously subgraph structures are extracted at each node of a decision tree by stepwise pair expansion in GBI to be used as attributes for testing. A beam search is employed in GBI to extract good enough discriminative subgraphs within the greedy search framework. Pessimistic pruning is incorporated to avoid overfitting to the training data. Experiments using a DNA data set indicate that DT-GBI can construct decision trees with only little prior domain knowledge, while still retaining results comparable to other classifiers that did use additional domain knowledge. Other experiments using a real-world hepatitis data set reveil patterns matching the experience of doctors.
7. The seventh paper by Yun Chi et al. gives an overview of a broad range of tree mining algorithms and identifies common theoretical foundations. The algorithms are compared and categorized according to their problem definitions and the techniques to solve various subtasks of the subtree mining problem. Especially, the review focuses on two main components of the algorithms, *i.e.*, the candidate generation step and the support counting step. It also presents a thorough performance evaluation for a representative family of algorithms, and clarifies their computational characteristics. Finally, it discusses some important issues in tree mining, such as the relationship between multi-relational data mining and tree mining.

Editors

Takashi Washio
The Institute of Scientific
and Industrial Research,
Osaka University,
8-1, Mihogaoka, Ibaraki, Osaka,
567-0047, Japan

Joost N. Kok
Leiden Institute
of Advanced Computer Science,
Leiden University, Niels Bohrweg 1,
2333 CA, Leiden, The Netherlands

Luc De Raedt
Institut fur Informatik,
Albert-Ludwigs-University Freiburg,
Georges-Koehler-Allee, Building 079,
D-79110 Freiburg, Germany

References

[1] Quinlan, J.R.: C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers*, 1993.

[2] Agrawal R. and Srikant R. Fast algorithms for mining association rules, *Proceedings of 20th International Conference on Very Large Data Bases* (VLDB 1994), pp. 487–499, 1994.

[3] *ECML Workshop on Multi-Relational Data Mining*, MRDM 2001,
http://mrdm.dantec.nl/
1st KDD Workshop on Multi-Relational Data Mining, MRDM 2002,
http://www-ai.ijs.si/SasoDzeroski/MRDM2002/
2nd KDD Workshop on Multi-Relational Data Mining, MRDM 2003,
http://www-ai.ijs.si/SasoDzeroski/MRDM2003/
3rd KDD Workshop on Multi-Relational Data Mining, MRDM 2004,
http://www-ai.ijs.si/SasoDzeroski/MRDM2004/

[4] *14th International Conference on Inductive Logic Programming*, ILP 2004,
http://ilp.fe.up.pt/

[5] *1st International Workshop on Mining Graphs, Trees and Sequences*, MGTS 2003,
http://www.ar.sanken.osaka-u.ac.jp/MGTS-2003CFP.html
2nd International Workshop on Mining Graphs, Trees and Sequences, MGTS 2004,
http://hms.liacs.nl/mgts2004/