Graph-Based Concept Discovery

Alev Mutlu
Kocaeli University, Turkey

Pinar Karagoz
Middle East Technical University, Turkey

Yusuf Kavurucu
Turkish Naval Research Center Command, Turkey

INTRODUCTION

Multi-relational data mining (MRDM) (Džeroski, 2003) is concerned with discovering patterns hidden in data that is stored in relational format, i.e. as a database. One of the most commonly addressed tasks in MRDM is concept discovery (Džeroski, 2003), where the problem is inducing logical definitions of a specific relation, called *target relation*, in terms of other relations called *background knowledge*. Given a database consisting of kinship relations such that *mother* is the target relation; *father* and *wife* and are among background knowledge, a typical concept discovery system would induce a concept rule like *mother(A,B):-father(C,B), wife(A,C)*.

Concept discovery can be considered as a predictive learning task where target relation instances are generally represented as ground facts, background knowledge is represented either intensional or extensional; and concept descriptors are usually in the form of Horn clauses. Logic-based, more specifically Inductive Logic Programming (Muggleton, 1991) (ILP)-based, and graph-based approaches are two competing counterparts in concept discovery. ILP-based approaches (Quinlan, 1990; Muggleton 1995) benefit from the powerful data representation framework provided by first order logic and the easily interpretable concept descriptors they discover. Such systems suffer from long running times mainly due to evaluation of the large search space they build, and are vulnerable to miss certain concept descriptors due to the local minima problem (Richards & Mooney, 1992). Graph-based approaches (Gonzalez, Holder, & Cook, 2002; Yoshida & Motoda, 1995) also provide powerful mechanisms to represent relational data and have heavily studied algorithms that can be modified to find concepts in graph data. Graph-based concept discovery systems can be classified as substructure-based approaches and pathfinding-based approaches. In the former, the assumption is that concepts should appear as frequent substructures in a graph, hence computationally expensive algorithms like subgraph isomorphism need to be employed. The assumption behind the systems that fall into the second category is that concepts should appear as finite length paths that connect certain vertices, hence such system need to employ path-profiling algorithms.

In this chapter we aim to introduce concept discovery problem and discuss several issues of graph-based concept discovery in depth. To this aim, in this chapter, we provide fundamentals of concept discovery in general, and in detail discuss state-of-the-art graph-based concept discovery methods by means of (i) data representation, (ii) search method, and (iii) concept evaluation mechanism.

The rest of the article is organized as follows. In the second section, an introduction to concept discovery and definitions of fundamental terms in concept discovery are provided; the concept

DOI: 10.4018/978-1-5225-2255-3.ch171

discovery problem is formally expressed. In the third section, classification of graph-based concept discovery approaches is provided; knowledge representation, search methods, and evaluation methods are introduced. In this section a running example is provided for a state-of-the-art pathfinding-based system presented in (Abay, Mutlu, & Karagoz, 2015a). In the fourth section, possible research directions for graph-based concept discovery are presented.

BACKGROUND

Multi-relational data mining is concerned with inducing patterns hidden in data stored in relational model, i.e. in a relational database. Traditional data mining algorithms are designed to work on flat files and cannot directly be work on such data. Two main directions in mining relational data are (i) converting the relational data into single table and then employing propositional learners (Kramer, Lavrač, & Flach, 2001), and (ii) developing new algorithms that can directly work on relational data (Džeroski, 2003). The main disadvantage of the first approach is the possible loss of information during propositionalization process. Most of the algorithms that belong to the second direction have their roots in ILP. ILP is concerned with inducing general theories from examples. It was first proposed for learning binary classification rules, but now can model more complex data mining problems.

One of the most commonly addressed tasks in MRDM is concept discovery. Concept is a set of frequent patterns embedded in the features of the concept instances and its relations to other objects (Toprak, Senkul, Kavurucu, & Toroslu, 2007). *Completeness* and *consistency* are two measures to test quality of concept descriptors (Muggleton & De Raedt, 1994). A concept descriptor is called complete if it explains all of the positive target instances and consistent if it explains none of the negative target instances. As real world data is usually noisy, completeness and consistency are hard to meet in their pure definitions, hence they are extended to explain as many positive target instances as possible and as few negative target instances as possible, respectively (Pham & Afify, 2006). Concept discovery systems generally follow iterative covering approach to induce concept descriptors. They start with an initial hypothesis

Algorithm 1. Generic concept discovery algorithm

```
Input: E: target instances, B: background knowledge
Output: H: Complete and consistent concept descriptors
1: while E ≠ ∅ or H' ≠ ∅ do
2:         H' = ∅
3:         Start with an initial hypothesis set H'
4:         for all h ∈ H' do
5:                 refine(h)
6:                 evaluate(h, B)
7:                 if good(h) then
8:                         cover(E, h)
9:                         H'=H' ∪ h
10:                 end if
11:         end for
12:         H = H ∪ H'
13: end while
14: return H
```