Relational and Semantic Data Mining for Biomedical Research

Nada Lavrač, Petra Kralj Novak

Dragan Gamberger, Filip Železny, Igor Trajkovski, Vid Podpečan, Igor Mozetič, Kristina Gruden, Hannu Toivonen, Anže Vavpetič et al.
Relational data mining: propositionalization

Learning from multi-relational data
Overview

- Relational data mining: propositionalization
  - Learning from multi-relational data

- Semantic data mining
  - Learning by using domain ontologies as background knowledge
RSD = Relational Subgroup Discovery

Železny and Lavrač, ML 2007
RSD: Propositionalization approach to relational data mining

Relational representation of customers, orders and stores.
RSD: Propositionalization approach to relational data mining

Step 1

1. constructing relational features
2. constructing a propositional table

Relational representation of customers, orders and stores.
Relational subgroup discovery with RSD

Step 1
Propositionalization
1. constructing relational features
2. constructing a propositional table

Step 2
Subgroup discovery

patterns (set of rules)
Domain ontologies as background knowledge
SEGS: Search for Enriched Gene Sets

Trajkovski et al., JBI 2008
Gene Ontology (GO)

- GO is a database of terms, describing gene sets in terms of their
  - functions (12,093)
  - processes (1,812)
  - components (7,459)
- Terms are connected (is_a, part_of)
- Levels represent terms generality
- Genes are annotated to GO terms
Semantic subgroup discovery with SEGS
SegMine: SEGS + Biomine
Podpečan et al., BMC Bioinformatics 2011,
The Biomine graph contains information from public biomedical databases:

- **nodes (~1 mio)** correspond to different concepts (such as gene, protein, domain, phenotype, biological process, tissue)
- **edges (~7 mio)** connect related concepts

The Biomine query and graph exploration engine answers queries to potentially discover new links between entities.
SegMine: SEGS + Biomine

Microarray

gene1: ++
gene2: +
gene3: +
...
geneN: --

Gene sets

SEGS

Biomine

Exploratory link discovery

- - -
Semantic Data Mining:
SDM-SEGS and SDM-Aleph
Vavpetič and Lavrač, COMPJ, 2012
SDM task definition used in our work

Given:
- transaction data table, relational database, text documents, Web pages, …
- one or more domain ontologies

Find: a classification model, a set of patterns
- Discovers subgroups both for ranked and labeled data
- Exploits input ontologies in OWL format
Wrapper of the popular ILP system Aleph

Same inputs/outputs as SDM-SEGS

Any number of additional binary relations

1 Ashwin Srinivasan
http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html
Implementations

- Orange4WS (Podpečan et al., 2011)

- ClowdFlows browsed-based DM platform (Kranjc et al. 2012)
Summary: Relational and Semantic DM in Context

Data Mining

Relational Data Mining

Relational Subgroup Discovery

Semantic Web

Ontologies

Semantic Data Mining

Semantic Subgroup Discovery
Future work

- We may envision a paradigm shift from data mining to knowledge mining.
- The envisioned future Semantic data mining scenario in mining the Semantic Web:
  - mining knowledge encoded in domain ontologies,
  - constrained by annotated (empirical) data collections.