Mining Linked Open Data: a Case Study with Genes Responsible for Intellectual Disability

Gabin Personeni1, Simon Daget1, Céline Bonner2, Philippe Jonveaux3, Marie-Dominique Devignes1, Malika Smal-Tabbone1, Adrien Couliet1

1. LORIA (CNRS, Inria NGE, Université de Lorraine, FRANCE) 2. Laboratoire de Génétique Médicale, CHU de Nancy, FRANCE, INSERM U-954

Contact : gabin.personeni@loria.fr

Motivation

Increasing amounts of life sciences data are made available through the Linked Open Data (Bio2RDF, EMBL-EBI RDF platform). Challenge: use Linked Open Data to answer a biological question.

Approach:
1. query and integrate Linked Open Data,
2. mine the collected data using Inductive Logic Programming, a relational data mining method,
3. assess the contribution of domain knowledge in the Linked Open Data to the quality of both characterization and prediction

Application: characterize genes responsible for Intellectual Disability and build a model predicting whether a gene is responsible for Intellectual Disability or not.

Characterization of genes responsible for Intellectual Disability (ID).

Input:
- list of 282 genes responsible for ID (Inlow and Reisfleisch, 2004).
- list of 267 genes responsible for other phenotypes according to OMIM.
- data on these genes available in the Linked Open Data (LOD).

Experiments:
- can perform inference on domain knowledge (GO examples)

Principles:

Inductive Logic Programming (ILP):
- Input data on these genes available in the Linked Open Data (LOD).
- Involves the mapping of the relationship, plus the mapping of its domain and range
- One query per relationship: example with the gene_reaction relationship

Mapping of data to the Linked Open Data

SPARQL Queries

Data Integration

Triple Store

Data Mining

Theory
First-order logic rules

Entity-Relationship Model covering the data related to the biological question

- Restricted to binary relationships to fit the LOD
- Domain knowledge in the form of the Gene Ontology is-a hierarchy
- Query for relationship R retrieves couples (x,y) such that R(x,y)

Characterization:
- no-GO theory is highly specific: points to individual reactions and pathways, low amount of false positives, but low coverage of positive examples
- GO ones theories rarely use other data than GO annotations, but better coverage of positive examples

Prediction evaluation:
- Scores computed using leave-one-out cross-validation (KNIME workflow)
- Sensitivity and accuracy improves with each additional inference level

Experiment

Sensitivity (%)
Specificity (%)
Accuracy (%)

TP/P
TN/N
TP/(P+N)
TN/(P+N)

no-GO
28.6
94.1
59.6

GO1
47.9
81.3
64.1

GO2
55.7
80.5
67.8

GO3
55.7
81.7
68.3

GO4
52.1
83.1
68.6

Leave-one-out cross-validation evaluation for each experiment

References:


Rémi Enslen et al. Methodology and results for the extraction of knowledge from the BioPortal Logical Inductive (PLI) 1. 1st Conference Transparents on l’Extraction et la Gestion des Connaissances, Toulouse, 23 janvier 2013 (poster)