Towards a semantic repository of data mining and machine learning datasets

Ana Kostovska, Sašo, Džeroski, Panče Panov

SiKDD 2018:
Conference on Data Mining and Data Warehouses
Problem definition

- The amount of data that is being created and stored on a global level is enormous, and it still keeps growing.

- In science, progress is best achieved by reproducing and improving someone else's work.

- Data plays a central role in the process of knowledge discovery.

- Collecting high-quality data can be a time-consuming process and a huge investment.

- FAIR data (Findable, Accessible, Interoperable and Reusable)

- Our goal is to develop a semantic repository of ML datasets to improve the quality and performance of DM and ML by refining data management.

- Search and query datasets and metadata.
Semantic web technologies
SPARQL & Semantic Inference

@prefix OntoDM: <http://www.ontodm.com/clus/>

OntoDM:OntoDM_clus_00001 rdf:type "binary classification dataset"
OntoDM:OntoDM_clus_00002 rdf:type "binary classification dataset"
OntoDM:OntoDM_clus_00003 rdf:type "binary classification dataset"
OntoDM:OntoDM_clus_00004 rdf:type "multi-class classification dataset"
OntoDM:OntoDM_clus_00005 rdf:type "regression dataset"
OntoDM:OntoDM_clus_00006 rdf:type "multi-class classification dataset"

PREFIX OntoDM: http://www.ontodm.com/clus/>

SELECT ?x
WHERE
{
  ?x rdf:type "flat classification dataset"
}
Existing dataset repositories (UCI ML repository)

UCI ML repository:
- Statistical descriptors
- Source
- Dataset information
- Attribute information
- Relevant Papers
- Papers that cite this dataset
Existing dataset repositories (OpenML)

This dataset classifies people described by a set of attributes as good or bad credit risks. This dataset comes with a cost matrix:

<table>
<thead>
<tr>
<th></th>
<th>Good</th>
<th>Bad (predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>0</td>
<td>1 (actual)</td>
</tr>
<tr>
<td>Bad</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

It is worse to class a customer as good when they are bad (5), than it is to class a customer as bad when they are good (1).
Existing dataset repositories (Google Dataset Search)

- **Kaggle**
  - Parkinson's Disease Observations
  - Observations
  - www.kaggle.com
  - Updated Oct 2, 2016

- **PLOS**
  - Data from: Augmented frontal cortex diacylglycerol levels in Parkinson's...
  - figshare.com
  - Updated Mar 8, 2018

- **PLOS**
  - Data from: Baseline C-reactive protein level and life prognosis in Parkinson's...
  - datacyclopedia.org
  - Published Jul 31, 2015

- **Kaggle**
  - Parkinson's Vision-Based Pose Estimation Dataset
  - www.kaggle.com
  - Updated Aug 31, 2017

**kaggle**

- **Parkinson's Disease Observations**
- **Variables Regarding Parkinson's Disease**

- **Authors**
  - KrishnaThiyagarajan

- **License**
  - Unknown

- **Available download formats from providers**
  - CSV, ZIP

- **Description**
  - Gives a bunch of data regarding parkinson's disease. Some of these variables have a very high correlation with each other.
Ontologies and vocabularies for describing data and metadata

- Ontologies for describing machine learning and data mining:
  - OntoDM (OntoDT, OntoDM-core, OntoDM-KDD)
  - The Data Mining OPtimization Ontology (DMOP)

- Metadata vocabularies and ontologies for describing provenance information:
  - Dublin Core vocabulary
  - The PROV Ontology (PROV-O)
  - schema.org

- Domain related ontologies
Proposed solution – Architectural design I
Proposed solution – Architectural design II
Proposed solution – Architectural design III
Summary and further work

• Design of semantic repository

• Dataset annotations in PROBMOT

• Develop a user interface for the system

• Select the most suitable architecture
Thank you for your attention!

Questions?