Towards a semantic store of data mining models and experiments

Ilin Tolovski, Sašo Džeroski, Panče Panov

SiKDD 2018: Conference on Data Mining and Data Warehouses
Motivation

• Thousands of data mining experiments producing thousands of models are executed daily

• None, or very little information regarding the experimental setup, the algorithms used, their parameter settings, or the produced models are stored

• Dumping the results in a conventional storage database is not sufficient

• We lose time and resources in order to verify, reproduce, or improve the results

• There is a need for a system that will semantically annotate, and store these experiments, their results, and the produced models
Introduction

• Annotate: to make or furnish critical or explanatory notes or comment (Merriam-Webster)

• In CS terms, semantic annotation provides us with essential metadata

• Ontology: A set of concepts and categories in a subject area or domain that shows their properties and the relations between them

• Storing the metadata provides structure to the data and gives provenance information about it

• Enables us to perform automatic inference and extract new facts

• With this structure & metadata, we can:
  • Verify the data
  • Reproduce experiments
  • Improve the research
  • Semantically query the annotated models and experiments
Semantic annotation

- **Iris dataset**
- **Multi-class classification task**
- **Decision tree J48 algorithm**
- **ModelDTJ48**
- **Classification accuracy**

**Experiment #1**

- **Jane123** executed-by
- **3** has-classes
- **IG** heuristic
- **/** max-depth
- **10** int. node: leaves
- **8** has-value
- **0.89** runtime(s)

**DM-experiment**
Proposed solution

- Ontology-based framework for annotation, storage, and querying of data mining models and experiments

![Figure 1. Schema of the proposed solution](image-url)
Background

• Semantic Technologies overview
  • RDF
  • SPARQL
  • RDFS/OWL
  • NoSQL databases

• Scientific overview
  • Ontologies & vocabularies
  • Representation of ML models and experiments
  • Experiment and model databases
  • Annotation frameworks

Related work

• Ontology resources
  • OntoDM (Panov et al. 2008)
    • Unified framework for DM entities
    • Defines top level concepts in DM and ML
    • Represents complex data mining tasks
      • Structured output prediction
      • Semi-supervised learning
      • Online learning
  • DMOP (Keet et al. 2015)
    • Defines three core segments of a DM process: data mining task, algorithm, and data mining workflow
    • Automation of algorithm and model selection by meta-data analysis
Related work

• Repositories of models and experiments
  • OpenML (Vanschoren et al. 2014)
    • Online repository for storing DM experiments
    • Datasets, experiment runs, and ML tasks are also stored
    • Separate models are not stored
    • No semantic querying or inference from the stored metadata

• BioModels (Le Novere et al. 2006)
  • Online repository for storing biomedicine models
  • Most models are not annotated
  • Others are manually curated
  • Need for automatization of the curation process
Storage & querying

- Multiple storage options
  - MongoDB
  - Elasticsearch
- Separate storage for models and experiments
- Querying engine will run on SPARQL
- SPARQL 1.1 allows “reasoning on the fly”
- Enables automatic inference and extraction of new knowledge from the data
The benefits?

• Organized storage enabling us to reproduce, reuse, verify DM models and experiments

• Save (a lot of) time

• Automatic inference provides us with new facts about the produced models and experiments

• Execute complicated queries
  • Select all the models trained on data representing X, that used RMSE as an error function, used in a regression task of any type, and required <2h to be trained

• Knowledge base like this can be used for meta-learning for tasks that have not been addressed yet (SOP, Online learning)
Future work

• Designing the interface for the system
• Integration with current data mining software (CLUS, ProBMoT)

• Annotation of process-based models, predictive clustering trees (PCTs) in different tasks
  • Primitive output predictive modelling tasks
  • Structured output predictive modelling tasks
  • Semi-supervised learning tasks

• Make the system compatible for extracting data for meta-learning
Thank you!