MAP-i PhD Program

Thesis Proposal

Inductive Logic Programming for Big Data

Proposed by Rui Camacho and Vítor Santos Costa

Porto 17th January 2014

Context

Inductive Logic Programming (ILP) is a well known approach to Multi-Relational Data Mining. A central feature of ILP is the use of First Order Logic to encode both the data and the models. This representational feature enables ILP to handle easily structured data and to construct highly complex models. It is also possible for ILP to integrate in the model information originated from different sources and encoded in different formats. One last advantage of ILP is that, most often, the constructed models are comprehensible. All of the above lead to the use of ILP to a large number of complex problems in a wide range of domains.

A major problem, however, is that ILP systems may take a long time in analysinging large data-sets. This efficiency drawback may be explained by the following reasons. The search (hypothesis) spaces are often very large and the evaluation of each hypothesis, which involves theorem proving, may be time consuming. The search space is also highly redundant.

This thesis proposal addresses efficiency issues of ILP systems.

Research proposal

To be able to use ILP to analyse larger data-sets several approaches are proposed to be researched in the thesis.

An important issue to improve efficiency is to speedup the search of the hypothesis space. This can be done by reducing *redundancy*, and developing new search strategies, possibly stochastic search strategies.

The use of *parallelism* is also a very promising line of research in improving the efficiency of ILP systems.

To handle large amounts of data *incremental theory construction*, using a part of the examples each time (like windowing) can be very useful.

A closer connection with the *Prolog engine* could improve example evaluation. Using caching and tabling can substantially speedup example's evaluation. Analysis of the proof trees of the examples is another interesting line of research to improve efficiency on the example's evaluation. Prolog optimisation of sequences of ILP frequent procedures (such as query packs) may also produce considerable speedups.

A close connection with *data-bases*, transferring operations to the DB management system may also lead to significant improvements.

Significant improvements in the research topics pointed out above will enable the application of ILP to much larger data sets.

Working local proposal

The work will be supervised by Rui Camacho (FEUP) and Vítor Santos Costa (FCUP).

Laboratory where the work will be done

The PhD work will be done in LIAAD & CRACS at INESC-TEC and Faculiesy of Engineering and Sciences of UP

External Researcher

Ashwin Srinivasan has an extensive and most relevant work in ILP. A. Srinivasan is also an Adjunct Professor at the School of Computer Science and Engineering, University of New South Wales; a Visiting Professor at the Computing Laboratory, Oxford, UK.