Imagine a biology student who needs to analyze and evaluate data measured in a lab. For her, as a domain expert in her field, the business and the data understanding phases of the data mining process are not a problem. The main challenge for her is to pre-process and analyze the data and gain useful knowledge from it.

Our work is aimed at helping people to analyze their data in a simple and user friendly way with no previous knowledge of data analysis nor data-mining.

Introduction

We designed and implemented a web based application which significantly simplifies data-mining processes.

Discovered challenges:
- Data understanding
  - How to automatically find out the character of data?
- Data preparation
  - How to find out which attributes are important and which are not?
- Modelling
  - How to choose the correct data-mining model?
  - How to choose the best hyperparameters for that model?
  - How to get these results within a few seconds instead of hours or days?

Our proposal:
- Automatic conversational system pre-processes the input data and generates questions for user to determine the further steps.
- Hundreds of differently preprocessed data files using computational cluster for fast and reliable results.
- Custom Meta-learning algorithm with Landmarking features speeds up the combined model and hyperparameter selection.

Discussion

We focused on classification problems in our current working prototype.

DEMO: http://s.ics.upjs.sk/~sbocko/winston

We will support regression and pattern mining techniques soon. Planned public release of this software is on 1st. of August 2015. Using the Meta-learning we were able to speed up the model recommendation time. For this purpose we combined our Landmarking approach with statistical characteristic approach presented by R. Neruda et al.

Results of our hyper-parameter search approach

Pros and Cons of existing solutions applicable to this problem

References


Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice