Web based data-mining assistant

Štefan Bocko, Tomáš Horváth
Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice

Introduction

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**.

Problem

How to guide a **non-technical** user through the whole data-mining process?

The solution should be:

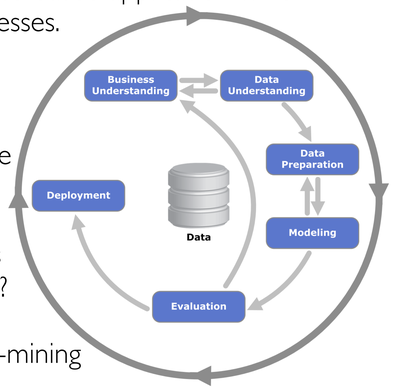
- user friendly
- easy to use
- fast (no long waitings for results)
- accurate

	bigml	OpenML	WEKA	rapidminer
No installation	✓	✓		
Easy-to-use	++++	+++	+	+
No technical skills required	✓	✓*		
Accurate results			✓	✓
Free		✓	✓	✓

↑ Pros and Cons of existing solutions applicable to this problem

Our approach

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



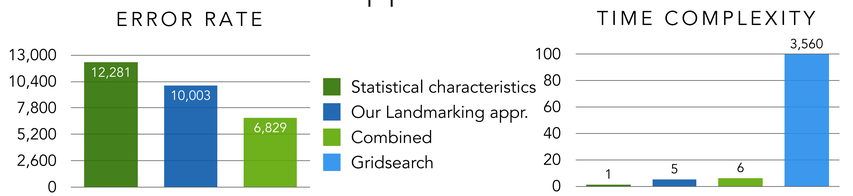
Discovered challenges:

- 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.

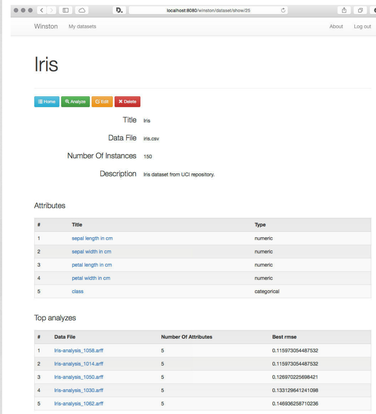
Results of our hyper-parameter search approach



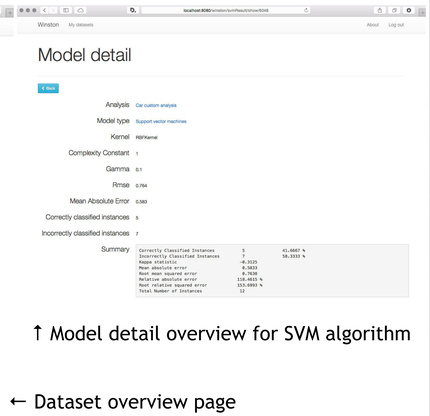
↑ Sum of positions of recommended models for different recommendation techniques on 52 datasets (183 484 models total).

↑ Number of trained models for different model selection algorithms.

More Results



← Dataset overview page



↑ Model detail overview for SVM algorithm

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.

References

1. Kazík, O., Pešková, K., Pilát, M., Neruda, R. Combining parameter space search and meta-learning for data-dependent computational agent recommendation. 11th International Conference on Machine Learning and Applications (ICMLA 2012): Boca Raton, Florida, USA, 12-15 December 2012. 2 volumes. ISBN 978146734651 1
2. Berka, P. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, 366 s. ISBN 80-200-1062-9.
3. Vilalta, R., Giruard-CARRIER, C., BRAZDIL, P. SOARES, C. Using Meta-Learning to Support Data Mining. International Journal of Computer Science & Applications, Vol. I, No. 1, p. 31–45. 2004.

stefan.bocko@student.upjs.sk