MAPi Thesis topic:

Development of a support system for workflow design for data mining problems that exploits metalearning

Context and Objectives

The aim of this thesis project is to develop methodologies useful when designing workflows for data mining problems. Workflows are now commonly used in many data mining systems, including Weka, Knime, Rapid Miner etc. Usually it is not easy to compose such workflows, as many possibilities exist. For instance, in classification different classifiers provide solutions of different quality. Some achieve higher accuracy rates than others. Different classifiers incur different costs related to the time spent in training. The solutions are affected by selection of certain variables and pre-processing. Similar issues arise when considering other data mining tasks (e.g. regression, optimization tasks etc.).

Thus there is a need for meta-level systems that help to organize and explore this vast space of possibilities. As has been shown in recent work, the design of workflows can be facilitated by recourse to planning systems. The resulting plan / workflow should provide a good compromise between solution quality and costs.

The aim of this project is to exploit the knowledge of what happened in the past (often referred to as metaknowledge), when deciding how to proceed in a new situation. This involves identifying promising candidate plans / workflows to adapt, or else, initiating a process of plan / workflow construction from scratch.

The candidate workflows need to be tested. The aim of these tests is to verify whether the goal of data mining has been achieved and determine the solution quality. The results of tests condition, in general, which candidate workflows should be considered next. The aim is to avoid generating and testing all possible workflows, which would incur high costs. The method should focus on "*promising workflows*" and *"promising experiments",* that is those experiments that have a high probability of providing new information, useful in the process of making the best decision. The solution should seek a good compromise between exploration and exploitation. Instead of carrying out all experiments from scratch, the work could exploit the existing experimental databases (e.g. the one at U. Leuven).

This thesis project could thus help to solve an important task of data mining which is how to design workflows for data mining problems in a more effective manner.

References

J.U.Kietz, F.Serban, A.Bernstein, S.Fischer: Designing KDD-Workflows via HTN-Planning, Proc. of PlanLearn 2012, associated with ECAI-2012.

Rui Leite, Pavel Brazdil: Active Testing Strategy to Predict the Best Classification Algorithm via Sampling and Metalearning, ECAI 2010, Proceedings of the 19th European Conference on Artificial Intelligence, Volume 215, page 309--314

Rui Leite, Pavel Brazdil, An Iterative Process for Building Learning Curves and Predicting Relative Performance of Classifiers, *Progress In Artificial Intelligence, Proceedings of the 13th Portuguese Conference on Artificial Intelligence Workshops (EPIA 2007)*, Volume 4874, page 87-98, 2007

Pavel Brazdil, Christophe Giraud-Carrier, Carlos Soares, R. Vilalta, *Meta-Learning: Applications to Data Mining,* Springer, 2009

Pavel Brazdil, Carlos Soares, Joaquim P. Costa, Ranking Learning Algorithms: Using IBL and Meta-Learning on Accuracy and Time Results, *Machine Learning, Volume 50, Number 3,* page 251-277, March 2003

Mahajan, A., Teneketzis, D.: Multi-armed bandit problems. In: Castanon, et al. (eds.) *Foundations and Applications of Sensor Management*. Springer-Verlag, 2007.

Fedorov, V.: *Theory of Optimal Experiments.* Academic Press, New York, 1972