MAP-I PhD Proposal Title: Efficient Multi-Objective Data Mining Optimization

Motivation:

Advances in information technologies have made it possible to collect, store and process massive, often highly complex datasets. All this data hold valuable information such as trends and patterns, which can be used to improve decision making. **Data mining (DM)** techniques aim at extracting high-level knowledge from raw data [1].

Two crucial issues in DM are [2,3]: 1) how to perform model and variable selection; and 2) how to evaluate the quality of a candidate model. The question of the first issue is: given a DM goal (e.g. classification or regression) what is the best set of input variables (i.e. features) and what is the best DM model? Currently, there are several feature selection (e.g. backward selection) and DM methods (e.g. decision trees, neural networks and support vector machines), each one with its own capabilities. Furthermore, several DM models can be aggregated, in what is known as an ensemble, and often ensembles provide more accurate predictions than individual DM models [4]. Turning to the second issue, business problems often include multiple quality criteria (objectives) to be optimized [5]. Thus, several **multi-objective optimization DM** methods have been proposed [3]. Among these, modern optimization techniques [5] (e.g. evolutionary computation) are particularly useful automatic search tools. In the literature, several works have proposed such techniques to optimize DM models (e.g. [6]). Yet, often these techniques require a substantial computational effort.

Objectives:

In this PhD thesis, we intend to study efficient optimization techniques for multi-objective DM. Such techniques may include evolutionary computation, particle swarms or hyperheuristics (i.e. build systems which can handle classes of problems rather than solving just one problem). In particular, the computational effort (e.g. number of searches or time) will be considered as an important dimension of this multi-objective optimization DM process. For instance, the developed automatic optimization method should be capable of providing the best DM model (according to several criteria) under a given time limit. And that the best provided model could evolve, depending on the computational effort that is available.

Supervisor:

Paulo Cortez, pcortez@dsi.uminho.pt, Department of Information Systems/Algoritmi R&D Centre, University of Minho, Guimarães, Portugal (http://www3.dsi.uminho.pt/pcortez)

References:

[1] E. Turban, R. Sharda, J. Aronson, D. King, Business Intelligence, A Managerial Approach, Prentice-Hall, 2007.

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[3] A. Freitas, A critical review of multi-objective optimization in data mining: a position paper, ACM SIGKDD Explorations Newsletter, 6(2): 77—86, 2004.

[4] T. Dietterich, Ensemble methods in machine learning, In Multiple Classifier Systems, Lecture Notes in Computer Science 1857, J. Kittler and F. Roli, Eds.Springer-Verlag, 2000, pp. 1–15.

[5] Z. Michalewicz, M. Schmidt, M. Michalewicz and C. Chiriac, Adaptive Business Intelligence, Springer, 2007.

[6] M. Rocha, P. Cortez and J. Neves. Evolution of Neural Networks for Classification and Regression. In Neurocomputing, Elsevier, 70 (16-18):2809-2816, October, 2007.