Universidade do Minho

MAP-I PhD Thesis Proposal

Title: Efficient Multi-Objective Data Mining Optimization

Thematic Area: Adaptive Business Intelligence

Introduction: Due to advances in Information Technology, nowadays it is easy to collect, store and process data. Vast datasets are becoming commonplace and all these data hold valuable trends and patterns, which can be used to improve decision making. Thus, an increasing emphasis is given towards the field of Data Mining, which aims at the extraction of useful knowledge from raw data. Under slight distinct perspectives, Data Mining is also known as (or strongly related with) Analytics, Business Intelligence, Data Science and Big Data. There are numerous examples of Data Mining successful applications. For instance, DM can be used to reduce fraudulent credit card use, to identify customer-buying habits and to predict monthly sales.

In this project, we focus on supervised learning, which includes two important DM goals: classification and regression. In both goals, a data-driven model is built, in order to model an unknown underlying function that maps several input variables into a desired target (i.e. the predicted dependent variable). There are several supervised DM algorithms, each one with its own advantages. In this project, we will address flexible and modern nonlinear learning methods, such as Neural Networks (NN) and Support Vector Machines (SVM), that often achieve high predictive performances when compared with more simpler methods (e.g. multiple regression) [2]. When applying these DM methods, variable and model selection are critical issues [3, 4]. Variable selection is useful to discard irrelevant inputs, leading to simpler models that are easier to interpret and that usually give better performances. Complex models may overfit the data, losing the capability to generalize, while a model that is too simple will present limited learning capabilities. Indeed, both NN and SVM have hyperparameters that need to be adjusted, such as the number of NN hidden nodes or the SVM kernel parameter, in order to get good predictive accuracy. 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 [5]. On the other hand, business problems often include multiple quality criteria (objectives) to be optimized (e.g. maximizing accuracy and minimizing the cost of the input attributes used by the data-driven model).

To solve these issues (variable and model selection, multiple quality criteria), several multi-objective optimization DM methods have been proposed [4][6]. Among these, modern optimization techniques [7] (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. [8, 9]). Since these methods already use a population of distinct models, it is easy to built ensembles without any extra computation cost. Yet, often these techniques require a substantial computational effort and do not consider computation as a scarce resource, since typically hundreds or thousands of models need to be fit and the training of a single model may require minutes or even hours of computation. The problem of computational effort is heavily enhanced when dealing with big data.

Objectives: The overall goal of this PhD is to study new 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) [7][10]. 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 or ensemble of DM models (according to several criteria) under a given time limit. And that the best provided model/ensemble could evolve, depending on the computational effort that is available. The proposed techniques will be applied in several real-world domains, such as time series forecasting [9], wine quality estimation [3] or civil engineering [11]. The computer simulations will be carried out using open source tools (e.g. rminer package of the R tool [12]). As the main outcome of this PhD, it is expected the development of fast automatic tools for the design of DM models/ensembles, capable of high performances under several user defined criteria and these tools will be particularly valuable for non expert DM users.

Supervisor: Paulo Cortez (Associate Professor with Habilitation), pcortez@dsi.uminho.pt, http://www3.dsi.uminho.pt/pcortez, ALGORITMI Research Centre/Department of Information Systems, University of Minho, Guimarães, Portugal.

R&D Unit: ALGORITMI Research Centre (http://algoritmi.uminho.pt/), University of Minho, Portugal.

External Researcher (to join the PhD Monitoring Group): Bernardete Ribeiro, Departmento de Engenharia Informática, Universidade de Coimbra, bribeiro@dei.uc.pt, https://www.cisuc.uc.pt/people/show/2020

References:

[1] E. Turban, R. Sharda, J. Aronson, D. King, Business Intelligence, A Managerial Approach, Prentice-Hall, 2007.
[2] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, NY, USA, 2008.
[3] 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.
[4] A. Freitas, A critical review of multi-objective optimization in data mining: a position paper, ACM SIGKDD. Explorations Newsletter, 6(2): 77—86, 2004.
[5] 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.
[6] C. Coello Coello. A Comprehensive Survey of Evolutionary-Based Multiobjective Optimization Techniques. Knowledge and Information Systems. An International Journal, 1(3):269-308, August 1999
[7] Z. Michalewicz, M. Schmidt, M. Michalewicz and C. Chiriac, Adaptive Business Intelligence, Springer, 2007.
[8] 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.
[9] J. Peralta, Xiaodong Li, G. Gutierrez, A. Sanchis. Time series forecasting by evolving artificial neural networks using genetic algorithms and differential evolution. In proceedings of International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, July 2010.
[10] E. Burke, G. Kendall, J. Newall, E. Hart, P. Ross and S. Schulenburg, Hyper-heuristics: An emerging direction in modern search technology, Handbook of metaheuristics, pp. 457-474, Springer, 2003.
[11] J. Tinoco, A. Correia and P. Cortez. Application of Data Mining Techniques in the Estimation of the Uniaxial Compressive Strength of Jet Grouting Columns over Time. In Construction and Building Materials, Elsevier, 25(3):1257-1262, March 2011.
[12] P. Cortez. Data Mining with Neural Networks and Support Vector Machines Using the R/rminer Tool. In P. Perner (Ed.), Advances in Data Mining - Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining (ICDM 2010), LNCS 6171, pp. 572-583, Berlin, Germany, Springer, 2010.