ROBUST DATAMINING

Uwe Aickelin

The University of Nottingham Ningbo China
Corresponding email: uwe.aickelin@nottingham.edu.cn

Abstract

Our long-term research goal is to develop datamining methodologies that are robust to changes in data and uncertainty. By robust we mean solutions remain 'optimal' when things change or are easily repaired. Broadly, this robustness can be achieved in two ways: One, by having 'slack' in the solution or two, by constructing the solution such that is easily repairable, e.g. failures are isolated.

Uncertainty in datamining can be introduced in many ways. Some of it can be due to unreliable data collecting, noisy data or simply continuous real-time and changing data streams. However, the part of uncertainty most of interest to us is that introduced by the human angle. For instance, we know from past research that the same experts make different decision based on the same data when approached a month later (Miller et al 2016). We also hypothesise that under certain conditions people change their behaviour or strategies, e.g. from co-operating to competing Fatah et al 2016).

In the field of optimisation, robustness has previously been explored extensively and there are some mature approaches such as stochastic programming (Bertsimas and Sim, 2004. In the field of datamining, this is a newer concept and only some basic approaches exist, like robust Principal Component Analysis (Xanthopoulos et al 2012). A completely novel approach could be a semi-supervised 'uncertainty coefficient' algorithm.

Part of the new methodology to solve this problem is to arrive at some new definitions. What do we mean by robust in a datamining context, e.g. what is the equivalent to 'slack' and 'reparability' in optimisation. Moreover, we could introduce an 'uncertainty coefficient' for input attributes. Could these coefficients work akin to 'privileged information' in Support Vector Machine approaches (Feyereisl and Aickelin, 2012)? In other words, they are neither input parameters nor labels, but semi-labels or rules that describe (some) of the data, but help the datamining in a semi-supervised way.

Therefore a good research project would be four work packages: First to establish that changing behaviour exists in classification data sets, second to obtain suitable definitions of 'robustness' (or 'slack' etc.), third to implement some established Operational Research or optimisation methods (such as stochastic programming or suitable metaheuristics) and try to use these to address the problem, fourth compare these to some novel but basic robust datamining methods (such as 'robust support vector machines') and then fifth implement a new advanced method and see how it rates. This paper will present the first two steps based on data collected from public good games.

References

[1] Petros Xanthopoulos, Panos Pardalos, Theodore Trafalis; Robust Datamining, Springer, 2012.
[2] Jan Feyereisl, Uwe Aickelin, Privileged information for data clustering, Information Sciences 194, 4-23, 2012.
[3] Simon Miller, Christian Wagner, Uwe Aickelin, Jonathan Garibaldi, Modelling cyber-security experts' decision making processes using aggregation operators, Computers & Security 62, 229-245, 2016.
[4] Polla Fattah, Uwe Aickelin, Christian Wagner, Optimising Rule-Based Classification in Temporal Data, ZANCO Journal of Pure and Applied Sciences 28 (2), 135-146, 2016.
[5] Dimitris Bertsimas, Melvyn Sim, the Price of Robustness, Operations Research. 52 (1): 35–53, 2004.