Recent advances in multivariate filter methods of variable selection for discrimination

Pädraic Walsh[1], Marta Garcia-Finana[2] and Gabriela Czanner[3]

[1] *University of Liverpool, United Kingdom, Padraic.Walsh@liverpool.ac.uk*
[2] *University of Liverpool, United Kingdom*
[3] *University of Liverpool, United Kingdom*

Keywords: variable selection for discrimination, probability of correct classification, filter, embedded

Variable selection is a common problem in discrimination when many potential predictors are considered. Filter methods are computationally fast and classifier-independent utilising a metric of discriminating ability to select variables however they may impose invalid assumptions on data. Embedded methods have high computational requirements (e.g. Random Forest). Hotelling's T^2 statistic is a multivariate index of the discriminating potential used by filter methods to select variables. It assumes equality of variance-covariance matrices across groups, which is often violated. We generalised Hotelling's T^2 statistic into the SNR ratio allowing heterogeneity of variance-covariance matrices across groups. We implemented SNR into a forward selection algorithm producing a novel method for variable selection. Using simulated data we demonstrated that SNR is better than T^2 at choosing the relevant discriminating variables (100 % vs. 46 %). In a comparison study with existing filter and embedded methods our algorithm demonstrated superior performance to filter methods and comparable performance to embedded methods but with reduced computational requirements. We investigated our methods in two clinical datasets: diabetic retinopathy (27 variables, and 103 patients) and a screening programme for sight-threatening diabetic retinopathy (17 variables, 5272 patients). We found that the variables chosen by SNR lead to better or equivalent classification accuracy compared to T^2 in terms of probability of correct classification (83 % vs. 76 %, and 76.5 % vs 76.5 %, respectively). We studied the performance of our methods for non-normal data; and demonstrated that our method is either superior or at least as good as alternative methods with high computational requirements. In our talk we will i) summarise recent advances in filter methods of variable selection for discrimination, ii) present our novel SNR describing its methodological properties in simulations, iii) present the forward selection algorithm discussing possible stopping criteria and iv) outline the challenges in applying SNR to real datasets.