The Application of Consensus Modelling and Genetic Algorithms to Interpretable Discriminant Analysis

Nathan Brown,[†] Milan Ganguly,[‡] Ansgar Schuffenhauer,[†] Peter Ertl,[†] Valerie J. Gillet,[‡] and Paulette A. Greenidge[†]

[†] Novartis Institutes for BioMedical Research, Basel, CH-4002, Switzerland and [‡] The Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom.

An evolutionary statistical learning method was applied as a discriminant analysis technique to classify drugs according to their biological target, and also to discriminate between a compilation of oral and non-oral drugs. The emphasis is placed not only on how well the models predict, but also on the interpretability of these models.

In an enhancement to previous studies, the consistency of the model weights over several runs of the genetic algorithm was considered in order to produce comprehensible models. Using this strategy, it was possible to identify the descriptors and their ranges that contribute most to class discrimination. Selecting a bin step size that enables the average descriptor properties of the class being trained to be captured improves the interpretability and discriminatory power of a model. The performance, consistency and robustness of such models were further enhanced by using two novel approaches that reduce the variability between individual solutions: consensus and splice modeling.

To gauge the quality of these consensus models, a set of comparative studies are presented using similarity searching, naïve Bayesian classifiers and support vector machines.