Statistical Sciences
UNIVERSITY OF TORONTO

SEMINAR

June 18, 2014 at 3:00pm

Sidney Smith Hall, Room 1074

Ensembling Classification Models Based on Phalanxes of Variables with Applications in Drug Discovery

Jabed H Tomal, Post-doctoral Fellow, University of British Columbia

Statistical detection of a rare class of objects in a two-class classification problem can pose several challenges. Because the class of interest is rare in the training data, there is relatively little information in the known class response labels for model building. At the same time the available explanatory variables are often moderately high dimensional. For instance, in our drug-discovery application, compounds are active or not against a specific biological target, such as lung cancer tumor cells, and active compounds are rare. Several sets of chemical descriptor variables from computational chemistry are available to classify the active versus inactive class; each can have up to thousands of variables characterizing molecular structure of the compounds. The statistical challenge is to make use of the richness of the explanatory variables in the presence of scant response information.

Our algorithm divides the explanatory variables into subsets data adaptively and passes each subset to a base classifier. The various base classifiers are then ensembled to produce one model to rank new objects by their estimated probabilities of belonging to the rare class of interest. The essence of the algorithm is to choose the subsets such that variables in the same group work well together; we call such groups phalanxes. The method is illustrated on four biological assays, each with five descriptor sets.