Abstract

Charles University in Prague
Faculty of Pharmacy in Hradec Králové
Department of Biophysics and Physical Chemistry

Candidate: Savvas Athanasiadis
Supervisor: Jurjen Duintjer Tebbens
Title of diploma thesis: The small sample size problem in gene expression tasks

The thesis addresses classification of genes to tumor types based on their gene expression signatures. The number of variables (amino-acids) to be investigated is typically very high (in the thousands) while it is expensive and time-consuming to analyze a high number of genes; usually at most tens of them are available. The combination of a small sample size with a large number of variables makes standard statistical classification methods inappropriate.

The thesis focuses on a modification of a standard classification method, Fisher’s linear discriminant analysis, for the case where the number of samples is smaller than the number of variables. It proposes an improved strategy to test this modified method with leave-one-out cross validation. Using so-called low rank updates of the involved covariance matrices, the computational costs of the cross validation process can be reduced by an order of magnitude. Memory demands are reduced as well.