Modeling Mutagenicity Status of a Diverse Set of Chemical Compounds by Envelope Methods

Abstract

Envelope models provide a flexible setup for dimension reduction in multivariate data analysis. They can work in conjunction with multivariate linear regression to produce estimates with reduced variance, and envelope-reduced data can also be used for class prediction by linear discriminant analysis. In this project we apply envelope methods on a dataset consisting of 508 diverse chemicals for two different purposes: estimating effects of the 307 predictors on mutagenicity of compounds, and employing discriminant analysis for the purpose of mutagenicity prediction. 'Two-deep' leave-one-out cross-validation is used for the purpose of prediction, meaning that envelope dimension of the predictive model is selected separately for each holdout compound. Rank deficiency in the data is tackled in two ways: by applying envelope methods on the first few principal components of the predictor matrix, as well as through introducing a ridge regression-like regularization parameter. Finally predictive performance of envelope models are compared with those of two earlier papers on the same dataset.

Keywords

Envelope models; Dimension reduction; Two-deep cross-validation; Chemometrics; QSAR