DOCTORAL THESIS

Kriging model approach to modeling study on relationship between molecular quantitative structures and chemical properties

Yin, Hong

Date of Award:
2005

Kriging Model Approach to Modeling Study on Relationship between Molecular Quantitative Structures and Chemical Properties

YIN Hong

A thesis submitted in partial fulfillment of the requirments

for the degree of

Doctor of Philosophy

Principal Supervisor: Prof. FANG Kai Tai

Hong Kong Baptist University

April 2005

Abstract

The molecular descriptors include various topological indices, quantum chemical descriptors, physicochemical parameters and so on. They all give structure descriptions of chemical compounds. Chemometrics, especially, quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) attempt to correlate physical, chemical and biological activities or properties with structural descriptors of compounds and find a suitable model, called metamodel, to establish relationships between molecule descriptors and activities or properties. The results are useful in theoretical and computational chemistry, biochemistry, pharmacology and environment research.

Techniques in multivariate analysis and data mining, such as ordinary least squares regression, principal components regression, partial least squares regression, multivariate adaptive regression splines and multivariate additive regression tree, are useful tools for modeling. Metamodels generated by these methods, basically are linear models with independently identical distributed (i.i.d.) random errors. However, the assumption of independent and identical distributed errors in general metamodel is not always true. For instance, many examples show that there can still be unacceptably large residuals compared to measurement errors in many models of QSAR/QSPR research. The reason for this may be diverse. The simplest and the most natural reflection on our mind is that the unaccepted residuals could be dependent. These dependent errors will present more information than independent situation. For instance, we might use a stationary Gaussian process $\{z(\mathbf{x}_i), i = 1, 2, \cdots, n\}$ instead of independent random variables ϵ_i's. In fact, the general Kriging approach just consists of parametric item and a stochastic process. In this thesis, we compared the Kriging models with other metamodels. Experiments showed that the proposed Kriging

approach could improve the regression models used widely in Chemometrics.

It is known that Kriging is an interpolating predictor which would be very beneficial for the fitting of the training data, but is not very so good for the predictions of the testing data when the data were collected with random noise $\epsilon(\mathbf{x})$. So if we add a disturbing input $\epsilon(\mathbf{x})$ in the original Kriging, the new Kriging model called empirical Kriging in some literature will provide more accurate prediction for the noisy data than the Kriging model. Many authors have paid attention to the merit of non-interpolating Kriging model. One of purposes of this thesis is to apply the empirical Kriging model to quantitative structure-activity relationship (QSAR) research. We demonstrate in the case study that the empirical Kriging model can significantly improve the prediction accuracy of other metamodels, including the Kriging models.

Table of Contents