THÈSE.

Variable selection and neural networks for high-dimensional data analysis

Application in infrared spectroscopy and chemometrics

NABIL BENOUDJIT

NOVEMBRE 2003

Thèse présentée en vue de l'obtention du grade de docteur en sciences appliquées

Faculté des sciences appliquées

Université catholique de Louvain

Université catholique de Louvain
Faculté des Sciences Appliquées
Département d'électricité
Laboratoire de Microélectronique

Variable selection and neural networks for high-dimensional data analysis: application in infrared spectroscopy and chemometrics

Jury:

Prof. **A. De Herde** (président)	Université catholique de Louvain, Belgium
Prof. **M. Verleysen** (promoteur)	Université catholique de Louvain, Belgium
Prof. **V. Wertz**	Université catholique de Louvain, Belgium
Prof. **M. Meurens**	Université catholique de Louvain, Belgium
Dr. Ir. **F. Rossi**	Université Paris Dauphine, France
Prof. **M. Saerens**	Université catholique de Louvain, Belgium

Thèse présentée par
Nabil Benoudjit
en vue de l'obtention du grade de
Docteur en Sciences Appliquées

Novembre 2003

Contents

Université catholique de Louvain
Faculté des sciences appliquées
Département d'électricité
Laboratoire de micro-électronique (Machine Learning Group)

Place du Levant 3, 1348 Louvain-la-Neuve, Belgique
Tél. 32 (0) 10 47 25 40 – fax 32 (0) 10 47 25 98
E-mail benoudjit@dice.ucl.ac.be
http://www.dice.ucl.ac.be/mlg/

This thesis focuses particularly on the application of chemometrics in the field of analytical chemistry. Chemometrics (or multivariate analysis) consists in finding a relationship between two groups of variables, often called dependent and independent variables. In infrared spectroscopy for instance, chemometrics consists in the prediction of a quantitative variable (the obtention of which is delicate, requiring a chemical analysis and a qualified operator), such as the concentration of a component present in the studied product from spectral data measured on various wavelengths or wavenumbers (several hundreds, even several thousands). In this research we propose a methodology in the field of chemometrics to handle the spectrophotometric data which are often represented in high dimension. To handle these data, we first propose a new incremental method (step-by-step) for the selection of spectral data using linear and non-linear regression based on the combination of three principles: linear or non-linear regression, incremental procedure for the variable selection, and use of a validation set. This procedure allows on one hand to benefit from the advantages of non-linear methods to predict chemical data (there is often a non-linear relationship between dependent and independent variables), and on the other hand to avoid the overfitting phenomenon, one of the most crucial problems encountered with non-linear models. Secondly, we propose to improve the previous method by a judicious choice of the first selected variable, which has a very important influence on the final performances of the prediction. The idea is to use a measure of the mutual information between the independent and dependent variables to select the first one; then the previous incremental method (step-by-step) is used to select the next variables. The variable selected by mutual information can have a good interpretation from the spectrochemical point of view, and does not depend on the data distribution in the training and validation sets. On the contrary, the traditional chemometric linear methods such as PCR or PLSR produce new variables which do not have an obvious interpretation from the spectrochemical point of view. Four real-life datasets (wine, orange juice, milk powder and apples) are presented in order to show the efficiency and advantages of both proposed procedures compared to the traditional chemometric linear methods often used, such as MLR, PCR and PLSR.