Processing of NMR data of oils and fish oils samples for classification with SVM, ANN and Multivariate analysis.

Application of Multivariate Analysis, Support Vector Machines and Artificial Neural Network to the Processing of Nuclear Magnetic Resonance data of olive oil and fish oil samples for classification of geographic origin and discrimination between wild and farmed fish.

Raffaella Folgieri1, 2, Serge Rezzi3, 4, Saeed Masoum1, Christophe Malabat2, Mehdi Jalali-Heravi1, Douglas Neil Rutledge2, David E. Axelson5, Kàroly Hèberger3, Carlo Mariani6, Fabiano Reniero3, Claude Guillou2

(1) Department of Chemistry, Sharif University of Technology, P.O. Box 11365-9516, Tehran, Iran
(2) Laboratoire de Chimie Analytique, CNRS UMR 214 INRA/INSA, P.G., 16, rue Claude Bernard, 75005 Paris, France
(3) European Commission, Joint Research Centre, Institute for Health and Consumer Protection, Physical and Chemical Exposure Unit, BIEVABS T.P. 740, 21020 Ispra (VA), Italy
(4) Present address: BioAnalytical Science, Metabolomics and Biomarkers, Nestlè Research Center, P.O. Box 44, 1008 Lausanne 26, Switzerland
(5) MRT Consulting, 8 Wilmot St., Kingston, Ont., Canada K7L 4V1
(6) Stazione Sperimentale Oli e Grassi, Milano, Italy
(7) Dipartimento di Scienze dell’Informazione Università degli Studi di Milano, Milano, Italy

Motivation

Traceability and control of origin of food products are very important for the consumers and for the European enforcement laboratories. There is a need in developing analytical methods to ensure compliance with labelling, i.e. the control of geographical origin giving also support to the denominated protected origin (PDO) policy, and the determination of the genuineness of the product by the detection of eventual adulterations.

The combination of 1H NMR (Nuclear Magnetic Resonance) fingerprinting with multivariate methods provides an original approach to study the profile of these oils in relation with geographical origin of olive oil or for discrimination between wild or farm origin for fish like salmon.

Exp1. NMR data of fish oils

Concerning the experiment on fish oil, we used Support vector machines (SVMs) as a novel learning machine in the authentication of the origin of salmon. SVMs have the advantage of relying on a well-developed theory and have already proved to be successful in a number of practical applications. The method requires a very simple sample preparation of the fish oils extracted from the white muscle of salmon samples.

The data set was divided into three sets of training (74 samples), monitoring (45 samples) and validation sets (33 samples). To analyze data we choose the 1H NMR because of the simple sample preparation and because the resulting spectra contain a lot of information related to the structures of the molecules (useful in classifying the salmon fish oil).

NMR analysis of complex samples is accompanied by variations in peak position and peak shape not directly linked to the sample (due to temperature variation and inhomogeneities in the applied magnetic field and instrumental stability). The alignment of the signal has been performed.

Results

Applied to NMR data of fish oils, the SVM has been able to distinguish correctly between the wild and farmed salmon; however ca. 5% of the country of origins were misclassified.

Different kernels have been tested on these data, and the results showed that the RBF kernel is the most reasonable choice because of its simplicity and ability to model data of arbitrary complexity. Moreover RBF use less hyperparameters, which influence the complexity of model selection (the polynomial kernel, for instance, requires more parameters).

The high added value of olive oil makes its control an important goal for EU producers and consumers. There is a need in developing reliable analytical methods to ensure compliance with labelling, i.e. the control of geographical origin giving also support to the denominated protected origin (PDO) policy, and the determination of the genuineness of the product by the detection of eventual adulterations.

Exp2. NMR data of fish oils

The high added value of olive oil makes its control an important goal for EU producers and consumers. There is a need in developing reliable analytical methods to ensure compliance with labelling, i.e. the control of geographical origin giving also support to the denominated protected origin (PDO) policy, and the determination of the genuineness of the product by the detection of eventual adulterations.

Materials and methods

Olive oil samples were provided by Stazione Sperimentale Oli e Grassi (Milano, Italy). NMR spectra were measured in the step-flow mode using a Bruker (Rheinstetten, Germany) DRX-500 instrument operating at 500.13 MHz.

Multivariate (chemometric) techniques are able to filter out the most relevant information from a spectrum, e.g. for a classification. In the experiment on olive oil samples, the principal component analysis (PCA) was carried out on the ~12,000 variables (chemical shifts) and four data sets were defined prior to PCA: Linear discriminant analysis (LDA) of the first 50 PCs was applied for classification of olive oil samples according to the geographic origin and year of production. The data analysis has been carried out with and without outliers, as well. Variable selection for LDA was achieved using (i) the best five variables and (ii) an interactive forward stepwise manner.

Using LDA on the external validation sets the correct classification of olive oil varied between 47 and 75% (random selection), and between 35 and 92% (Kernard–Stone selection (KSS)) depending on geographic origin (country) and production years.

A similar success rate could be achieved using partial least squares discriminant analysis (PLS DA). The success rate can be considerably improved by using probabilistic neural networks (PNN). Correct classification by PNN varied between 58 and 100% on the external validation sets. Other chemometric techniques, such as multiple linear regression, or generalized partial correlation, did not give better results.

Conclusions

Support Vector Machines were used as a new class of classification algorithms in a validated method for the confirmation of wild and farmed salmon and their origins to reduce the possibility of fraud. This method seems to be the most suitable one, because a limited number of data points (support vectors) are needed for the classification. We believe that the power of this method partly depends on the analytical method used for the analysis of the fatty acids of the fish oils as the more informative data produced by an analytical method, the more useful they are for the classification. The combination of 1H NMR spectroscopy with SVMs has provided a novel method for the classification of salmon.

Contact

Claude Guillou/Raffaella Folgieri
European Commission • DG Joint Research Centre
JRCP
Tel. +39 0332 78 5678 • Fax +39 0332 78 9303
E-mail: Claude.Guillou@jrc.it