Joint Research Centre

www.jrc.cec.eu.int

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 farm fish.

Raffaella Folgieri[1,7], Serge Rezzi[3, 4], Saeed Masoum1, Christophe Malabat[2], Mehdi Jalali-Heravi1, Douglas Neil Rutledge[2], David E. Axelson[5], Kàroly Hèberger[3], Carlo Mariani[6], Fabiano Reniero[3] Claude Guillou[3]

(1) Department of Chemistry, Sharif University of Technology, P.O. Box 11365-9516, Tehran, Iran
(2) Laboratoire de Chimie Analytique, UMR 214 INRA/INA 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, BEVABS T.P. 740, 21020 Ispra (VA), Italy
(4) *Present address:* BioAnalytical Science, Metabonomics and Biomarkers, Nestlé Research Center, P.O. Box 44, 1000 Lausanne 26, Switzerland
(5) MRi Consulting, 8 Wilmot St., Kingston, Ont., Canada K7L 4V1
(6) Stazione Sperimentale Olii e Grassi, Milano, Italy
(7) Dipartimento di Scienze dell'Informazione-Università degli Studi di Milano, Milano, Italy

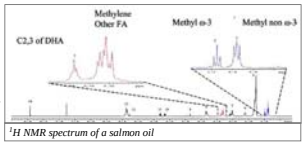
Motivations

Traceability and control of origin of food products are very important for the Consumers and for the European enforcement laboratories. There is thus a need in developing analytical methods to ensure compliance with labeling, i.e.the control of geographical origin giving also support to the denominated protected origin (DPO) 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 analysis 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 salmons.

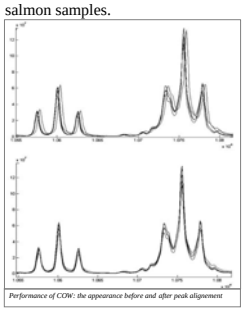
Exp1. NMR data of fish oils

Materials and methods

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.



^{1}H NMR spectrum of a salmon oil



Performance of COW: the appearance before and after peak alignement

The data set was divided into three sets of training (74 samples), monitoring (45 samples) and validation sets(22 samples)
To analyze data we choosed the ^{1}H 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 variation in peak position and peak shape not directly linked to the sample (due to temperature variation and inhomogeneities in the aopplied magnetic field and instumental stability). The alignement of the signal has been performed.

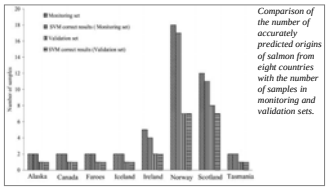
Results

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.

Sample	Given specification	Predicted wild (W)/farmed (F)	Predicted country of origin
1	Cofaws-blinds	F	Canada
2	Cofaws-blinds	F	Norway
3	Cofaws-blinds	F	Norway
4	Cofaws-blinds	W	Norway
5	Cofaws-blinds	F	Scotland
6	Cofaws-blinds	F	Norway
7	Cofaws-blinds	W	Norway
8	Cofaws-blinds	F	Canada
9	Cofaws-blinds	F	Norway
10	Cofaws-blinds	W	Ireland
11	Market-Canada-W	W	Ireland
12	Market-Canada-W	W	Ireland
13	Market-Canada-W	W	Ireland
14	Market-Canada-W	W	Norway
15	Market-Canada-W	W	Ireland
16	Market-Italy-F	F	Norway
17	Market-Italy-F	F	Norway
18	Market-Italy-F	F	Norway
19	Market-Italy-F	F	Scotland
20	Market-UK	W	Norway
21	Market-UK	F	Norway
22	Market-UK	F	Norway
23	Market-UK	F	Norway

Some classified results of predicion set for the two classification criteria using SVM.

Different kernels have been tested on these data, and the results showed that the RBF kernel is the most reasonable choice because of is 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)



Comparison of the number of accurately predicted origins of salmon from eight countries with the number of samples in monitoring and validation sets.

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 the data produced by an analytical method, the more useful they are for the classification. The combination of ^{1}H NMR spectroscopy with SVMs has provided a novel method for the classification of salmon.

Acknowledgements
The authors are grateful to the Europeanproject COFAWS (European Commission DG RTD FP5 project GRD2–2000–31813) and to all the collaborators from the partners of this project (Eurofins Scientific (Nantes- France), North Atlantic Fisheries College (Scalloway, Shetland Islands - United Kingdom), SINTEF Fisheries and Aquaculture (Trondheim-Norway), Joint Research Centre (Ispra-Italy)) who contributed to the collection and preparation of fish samples, and for the authorization to exploit their NMR data in this work.

EUROPEAN COMMISSION
DIRECTORATE GENERAL
Joint Research Centre

Exp2. NMR data of fish oils

The high added value of olive oil makes its control an important goal for EU producers and consumer. There is thus a meed in developing reliable analytial methods to ensure compliance with labeling, i.e. the control of geographical origin giving also support to the denominated protected origin (DPO) policy, and the determination of the genuineness of the product by the detection of eventual adulteration.

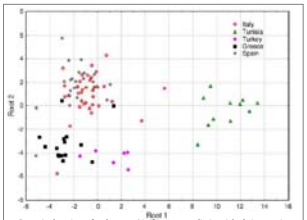
Materials and methods

Olive oil samples were provided by Stazione Sperimentale Olii e Grassi (Milano, Italy).
NMR spectra were measured in the stop-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 PC's 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.

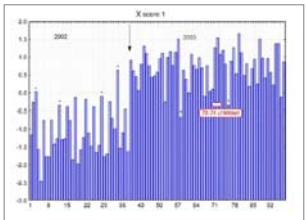
Data sets	Test set (random selection)	Methods				
		CC[a] (%)	GDA BS[b] (%)	LDA IS[c] (%)	PLS DA[d] (%)	PNN (%)
A1	Groups: 5, geographic origins: Italy 13, Tunisia 4, Turkey 2, Greece 5, Spain 8, irrespectively to the year of production, $n = 32$	20	53.2 (5)	65.6 (20)	62.5 (4)[g] 37.5 (4)[g]	61.0[h]
A2	Without outliers, groups: 5, geographic origins: Italy 10, Tunisia 4, Turkey 2, Greece 5, Spain 8, irrespectively to the year of production, $n = 29$	20	58.6 (5)	72.4 (20)	34.5 (1)[f,i] 37.9 (2)[g] 37.9 (4)[g]	60.0[h]
B	Groups: 5, geographic origins: Italy 1, Tunisia 3, Turkey 2, Greece 3, Spain 3, year 2002, $n = 12$	20	50.0 (5)[g]	75.0 (12)[g]	33.3 (1)[g,k]	58.3[h]
C	Groups: 4, geographic origins: Italy 10, Tunisia 0, Greece 2, Spain 5, year 2003, $n = 17$	25	82.4 (5)[f]	82.4 (12)[f]	58.8 (1)[g,i] 64.7 (3)[g]	76.5[h] 70.6[i]
D	Two groups: 20 samples from production year 2003 and 12 samples from 2002; irrespectively to the geographic origin, $n = 32$	50	65.6 (5)[g]	53.1 (19)[g]	80.0 (2)[g] 67.6 (2)[k]	84.0[h]

Analysis methods results comparison.

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% (Kennard–Stone selection (KS)) depending on geographic origin (country) and production years.



Canonical variates for data set A2 (five groups distinguished; interactive var. sel., 20 var. kept in the final model at the 5% level, random split).



PLS X score 1 for data set D: two groups differing in prod. year: 2003 and 2002, using two PLS comp. Sample nos. 70–74 from unknown prod. year.

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 pair-wise correlation, did not give better results.

1H NMR fingerprints acquired in high throughput mode hold appropriate information for classification of geographic origin and year of production.
Hundred percent correct classification can be achieved only on the training sets, but better classification is to be expected using designed experiments.
Classification by PLS shows peculiar characteristics. PLS DA is competitive if the number of groups is small. MLR is not recommended.
GPCM selects numerous variables, which are not significant at the 5% level.
PNN classification is a viable option for samples and methodologies of the type investigated here. Variability due to production year is less a factor than the actual country of original.

Contact

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

ihcp Institute for Health and Consumer Protection