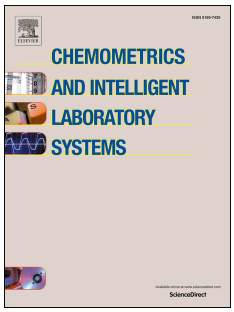
Accepted Manuscript



Artificially generated near-infrared spectral data for classification purposes

Vilma Sem, Jana Kolar, Lara Lusa

PII: S0169-7439(17)30232-0

DOI: 10.1016/j.chemolab.2017.11.009

Reference: CHEMOM 3542

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 7 April 2017

Revised Date: 8 September 2017

Accepted Date: 9 November 2017

Please cite this article as: V. Sem, J. Kolar, L. Lusa, Artificially generated near-infrared spectral data for classification purposes, *Chemometrics and Intelligent Laboratory Systems* (2017), doi: 10.1016/j.chemolab.2017.11.009.

Artificially generated near-infrared spectral data for classification purposes

Vilma Sem [a,*] Jana Kolar [b] Lara Lusa [c,d]

[a] Faculty of Agriculture and Life Sciences, University of Maribor, Pivola 10, 2311 Hoce, Slovenia
vilma.sem@um.si

[b] Kolar Morana RTD, d.o.o., Oslica 1b, 1295 Ivančna Gorica, Slovenia
jana.kolar@morana-rtd.com

[c] Department of Mathematics, University of Primorska, Glagoljaška 8, Koper, Slovenia
[d] Institute for Biostatistics and Medical Informatics, University of Ljubljana, Vrazov trg 2, Ljubljana, Slovenia
lara.lusa@mf.uni-lj.si

Abstract Near-Infrared Spectroscopy has became a widely used analytical technique in different research fields due to its non-destructiveness and low-cost. The spectra are rich in information but extremely complex, therefore their analysis necessitates the use of advanced statistical methods. The empirical properties of the statistical methods can be assessed using artificially generated data that resemble real Near-Infrared Spectroscopy.

In this paper we propose a new data generation approach (ABS) that takes into account the theoretical knowledge about the near-infrared absorption of the functional groups. The proposed method is compared to real data and to a simpler data generation method, which simulates the data from a multivariate normal distribution whose parameters are estimated from real data (MVNorig). The comparison between real data and the data generation approaches is based on a class-imbalanced classification problem using linear discriminant analysis, classification trees and support vector machines.

Both simulation approaches generated spectra with a good resemblance to real data, MVNorig performing slightly better than ABS; using real and simulated data we would have reached similar conclusions about the class-imbalance problem in classification.

Both methods can be used to artificially generate near-infrared spectra. The method based on multivariate normal distribution can be used when a large number of real data spectra are available, while the appropriateness of the results of the ABS method depend on the exactness of functional group near-infrared absorption knowledge.

Keywords: NIR spectroscopy, artificial data, simulation, classification analysis, class-imbalance.