Chemometrics and Intelligent Laboratory Systems 148 (2015) 72–76

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Design of a reference value-based sample-selection method and evaluation of its prediction capability

Zhonghai He *, Mengchao Li, Zhenhe Ma

College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning Province 110819, China

ARTICLE INFO

Article history:
Received 9 May 2015
Received in revised form 28 August 2015
Accepted 1 September 2015
Available online 8 September 2015

Keywords:
Reference value base
Sample selection
Prediction performance
Soy sauce samples

ABSTRACT

A calibration set comprises the multidimensional space that represents the samples for prediction. The representative ability of a calibration set is a major factor that affects the predictive performance of a multivariate regression. A new reference value (YR)-based sample-selection algorithm that assembles a dependent value (y-value) uniform distribution is presented to assure the representation. The existing typical sample-selection algorithm is used for comparison. A set of soy sauce data is used as a set of typical samples that have a complex solution. Comparing the prediction results, it is shown that YR sample-selection has similar prediction performance to that of sample set partitioning based on joint x–y distances (SPXY), but with a simpler algorithm. The calibration models of the y-reference-included sample sets (SPXY and YR) are more accurate than those of y-reference-excluded sample sets (RS and KS). After modeling with the selected representative samples, the performances of YR and SPXY are comparable to that of full sample modeling with fewer samples.

1. Introduction

In multivariate calibration problems, model performance is largely affected by the calibration samples that are used in the model-building process. A calibration set of n samples implies not only n number of spectra but also n reference measurements of a given response variable. A large n may lead to reliable models. However, in practical applications, n is usually rather small due to budget and/or time constraints related to the measurement of the response variable. Small calibration sets are usually prone to generate models with poor generalization abilities. A representative calibration set with limited samples can arrive at a better predictive capability. The calibration set sample-selection method is of value for analytical applications involving complex matrices, in which the compositional variability of real samples cannot be easily reproduced by optimized experimental designs of the entire set.

For online process measurement, many samples can be collected. However, eighty percent of the samples are similar or even duplicated, thus reducing the representation of the samples [1]. It is necessary to select a representative sample in model building. In real industrial processes, the concentration usually follows a normal distribution. If the samples are included in the calibration set without selection, the "Dunne effect", i.e., prediction results converge to a central value, will be induced, as shown in Fig. 1 [2]. Building a model using selected representative samples has the following advantages: an improved model building speed, a reduced storage requirement, and ease of updating the model.

Random sampling (RS) is a popular technique because it is easy to use. The Kennard–Stone (KS) algorithm is an improvement to RS that is often employed [3]. KS is aimed at covering the multidimensional space in a uniform manner by maximizing the Euclidean distances between the instrumental response vectors of the selected samples. However, a shortcoming of KS in the multivariate calibration context lies in the fact that the statistics of the dependent variable (y) are not taken into account. Sample set partitioning based on joint x–y distances (SPXY) [4] extends the KS algorithm by encompassing both the x- and y-differences in the calculation of the inter-sample distances.

To encompass the information of the distinct constituents, a dependent variable (y-value)-based sample-selection method (YR) is studied. The performances of YR, RS, KS, SPXY, and the full sample (FS) method are compared using the NIR spectra of sauce samples collected from a well-known sauce factory. The performances of the resulting models are compared in terms of the root-mean square errors calculated in the prediction sets not included in the modeling procedures. For the purpose of ensuring the independence of such sets, the prediction samples are randomly extracted from the initial pool of experimental data.

2. Materials and methods

2.1. Sample preparation

In total, 782 samples were collected randomly from a soy sauce factory with a collection procedure that lasted over one year. The division of the 782 samples into calibration and prediction sets was performed in

* Corresponding author. Tel.: +86 18603391927, +86 335 8064890.
E-mail address: professorhe@qq.com (Z. He).

http://dx.doi.org/10.1016/j.chemolab.2015.09.001
0169-7439/

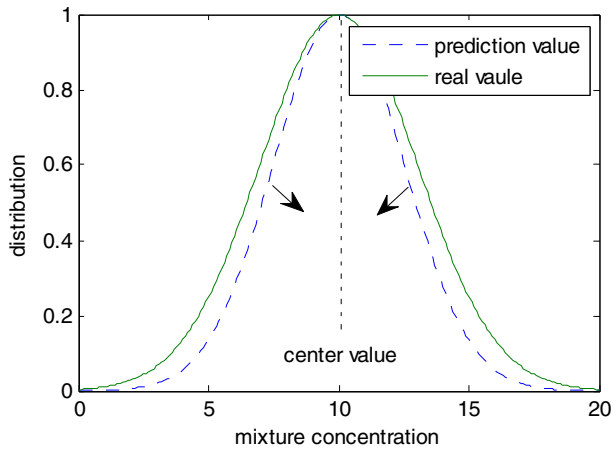

Fig. 1. Diagram of "Dunne effect".

the following manner: initially, 150 prediction samples were extracted from the full set in a random manner to simulate the analysis of a batch of real unknown samples, and the remaining 632 samples were used as a calibration set.

Three important ingredients in soy sauce were modeled individually, i.e., the ammonia nitrogen (AN), total acid (TA), and NaCl (CL), to illustrate the effects of representative samples on various calibration models. Reference analysis for TA, AN and CL was performed using the official methods for soy sauce according to the Chinese National Standard GB18186-2000. All ingredients were measured using a 1200-series high-performance liquid chromatography (HPLC) apparatus (Agilent Technologies, Inc., Santa Clara, CA, USA). The HPLC analysis was performed using a C18 column (150 × 4 mm, i.d., 5 μm).

The descriptive statistics for the AN, TA, and CL of the calibration and validation sets are presented in Table 1. The results revealed a broad range of parameters, especially in the calibration sets, which helps develop a stable and robust calibration model.

2.2. Spectra measurements

Spectral data were collected by measuring the diffuse reflectance from the soy sauce samples in the NIR region of 950–1650 nm at 2-nm increments using a diode-array analyzer (DA 7200, Perten Instruments, Sweden). Duplicates of each sample were scanned twice. The average spectrum of each sample was used in the subsequent data analysis. Because the sample is analyzed in an open dish, the problems associated with sample cups are avoided, and operator influence on the results is minimal. Typical spectra of soy sauce are shown in Fig. 2.

2.3. Spectral preprocessing

For almost all evaluation procedures, pretreatment of the data is required to extract useful information from the spectra. Baseline shifts in the spectral curve occurred for the raw NIR spectra of the soy sauce samples, and spectral preprocessing of the raw spectral data can be performed according to the usual method. In this study, the usual preprocessing method, i.e., the standard normal variate transformation (SNV) combined with the first-derived Savitzky–Golay method (1st derived, 9 points), is used [5]. The first derivative was adopted to sharpen the spectral profiles and eliminate the disturbances caused by potential baseline shift and background noise. SNV was used to remove slope variation and light scatter effects.

Table 1
Statistical results of the various sample sets.

Content	Calibration set (632 samples)				Prediction set (150 samples)			
	Min	Max	Mean	SD	Min	Max	Mean	SD
AN	0.3	0.88	0.59	0.074	0.4	0.95	0.554	0.097
TA	0.46	1.52	0.939	0.154	0.53	1.5	0.877	0.178
CL	14.72	22.18	19.04	0.7	17.84	20.39	19.3	0.552

3. Selection algorithms

3.1. General rule for calibration

The selection of calibration samples is crucial to the success of the quantitative analysis. In choosing the composition of the standards, the following points should be considered [6]:

(1) The calibration set should encompass the expected range of concentrations for the sample components. The sample set should be as extensive and well-distributed as possible [7]. Otherwise, a systematic prediction error will occur.
(2) Each solution should be a unique mixture. Redundant (equivalent or duplicate) standards prepared by diluting a stock solution provide no independent spectral information and undermine the quality of the calibration. This is the so-called "stock solution syndrome" [8].
(3) The sum of the component concentrations in the standards should not add up to a constant. This is known as the "100% syndrome". This causes problems during the mathematical operations of multicomponent analysis and may generate erroneous results [8].
(4) Calibration samples should be selected to ensure that the range of variation in the component concentrations is large relative to the precision to which the calibration samples can be made [9].
(5) The samples must be selected to have a distribution close to the uniform distribution [10,11].

3.2. Random selection

Random sampling (RS) is a popular technique because of its simplicity and also because a group of data randomly extracted from a large set follows the statistical distribution of the entire set. However, RS does not guarantee the representative nature of the set nor does it ensure that the samples on the boundaries of the set are included in the calibration.

3.3. KS selection algorithm

KS [3] has been widely used in quantitative spectroscopy and has shown good performance in terms of calibration sampling [12,13]. KS, initially called the uniform mapping algorithm, is a deterministic sequential approach that attempts to select samples uniformly distributed in the predictor space. The KS procedure to select a training or calibration subset of n samples ($X_{tr} = \{X_{trj}\}_{j=1}^{n}$) from a given set of N samples ($\mathbf{X} = \{x_i\}_{i=1}^{n}$, note that $n < N$) consists of the following:

(1) Find in $\mathbf{X}$ the sample x_{tr1} that is closest to the mean (μ), allocate it in x_{tr} and remove it from $\mathbf{X}$.
(2) Find in $\mathbf{X}$ the sample x_{tr2} that is the most dissimilar to x_{tr1}, allocate x_{tr2} in x_{tr} and remove it from $\mathbf{X}$.
(3) Find in $\mathbf{X}$ the sample x_{tr3} that is the most dissimilar to those already allocated in x_{tr}. Allocate x_{tr3} in x_{tr} and then remove it from $\mathbf{X}$. Note that the dissimilarity between x_{tr} and each x_i is given by the minimum distance of any sample allocated in x_{tr} to each x_i.
(4) Repeat step (3) $n-4$ times to select the remaining samples (x_{tr4}, … x_{trn}).

For distance computations in the KS algorithm, the Euclidean distance is commonly used.