Article

NIR news

Bypassing NIR pre-processing optimization with multiblock pre-processing ensemble approaches

NIR news
2022, Vol. 33(7–8) 5–8

DOI: 10.1177/09603360221139227
journals.sagepub.com/home/nir


Puneet Mishra

Abstract

Pre-processing near-infrared spectral data is a major part of near-infrared data modelling. A wide range of pre-processings are available to deal with both the additive and the multiplicative effects. However, practitioners have majorly focused on the selection of the best pre-processing technique or their combination. Data pre-processed with different pre-processings carry complementary information; hence, a natural solution to avoid pre-processing selection and to learn complementary information is the ensemble modelling. Recently, multiblock data fusion modelling-inspired ensemble techniques have gained momentum and several innovative approaches have been proposed for modelling near-infrared data. This article provides a state of the art of the new multiblock modelling-inspired pre-processing ensemble techniques. Their novelties and pitfalls are also discussed.

Keywords

Information fusion, ensemble learning, multiblock fusion, spectroscopy

Introduction

Near-infrared (NIR) spectroscopy is a widely used non-destructive technique for qualitative and quantitative analysis of materials. There are two building blocks for NIR spectroscopy, i.e., instrumentation and calibration modelling. The innovations in instrumentation for recording NIR spectra have almost reached saturation with recent developments in the miniaturisation of the instruments to have easy-to-operate portable devices. However, calibration modelling is still a topic of research where the scientific community is busy developing novel data modelling and model maintenance approaches. One such challenge is optimising NIR data pre-processing before the model development. NIR data intrinsically contain both light absorption and scattering due to the interaction of light with the chemical components and physical structure of materials. While modelling chemical components, it is beneficial to eliminate physical light scattering effects with spectral pre-processing approaches. However, the availability of many pre-processings requires the user to select the optimal pre-processing (or their combination) during the model development stage. Due to many pre-processings, the task of pre-processing selection is computationally expensive and often rely on the experience of the user. Furthermore, knowing that different pre-processings may eliminate information brings to the thought if selecting one pre-processing (or a single combination) is the best approach to NIR data modelling.

To avoid the selection of a single pre-processing and the deselection of the other candidate pre-processings, a new trend of ensemble pre-processing has emerged.[1] In ensemble pre-processing modelling, the aim is to use complementary information from differently pre-processed data. There are diverse ways of ensemble pre-processing modelling, e.g., one of the easiest pre-processing ensemble modelling approaches involves developing individual models for each pre-processing and then averaging the final predictions. However, there are better more interpretable ways to perform pre-processing ensembles based on multiblock data fusion modelling. Multiblock modelling[2] is a special field of chemometrics that deals with data fusion from multiple sources. The motivation behind multiblock methods is that they are highly interpretable and include methods which are data scale independent. Ensemble pre-processing can be considered a special case of data fusion where the complementary information from differently pre-processed data is modelled using multiblock methods.

Wageningen Food and Biobased Research, Wageningen, the Netherlands

Corresponding author:
Puneet Mishra, Wageningen Food and Biobased Research, Bornse Weilanden 9, P.O. Box 17, 6700AA Wageningen, the Netherlands.
Email: puneet.mishra@wur.nl

Multiblock-inspired pre-processing ensemble methods

Recently, three new multiblock-inspired pre-processing ensemble methods have been proposed for NIR data modelling. These methods are sequential pre-processing through orthogonalization (SPORT),[3] parallel pre-processing through orthogonalization (PORTO)[4] and pre-processing ensemble with response-oriented sequential alternation (PROSAC).[5] Note that methods are different in the way they model the differently pre-processed NIR data. For example, the first proposed method SPORT (Figure 1 (a)) is inherently a sequential method which is suitable when the user is aware of the order of pre-processings to learn the ensemble model. However, knowing the pre-defined order for pre-processing is not natural and one of the recommendations made by the developer of the method is to use easy and faster model free pre-processing at the start and computationally expensive pre-processing in later steps of sequential modelling. It is noteworthy that as the number of pre-processing blocks increases the information modelled from the later block gets scarce; hence, SPORT capability is limited in terms of modelling information from many differently pre-processing data. One of the main capabilities of SPORT is that being a sequential method, it models each data block individually; hence, SPORT is data scale independent and highly suitable for combining, for example, raw data with derivative pre-processed data, as usually the scale for them is very different. The limitation of SPORT to define the pre-processing order and capability to model low number of pre-processings led to the development of the second method called PORTO (Figure 1(b)). In PORTO, differently pre-processed data are modelled in the framework of extracting common and distinct information to explain the response. In PORTO, there is no need to define the pre-processing order, as it operates in parallel to differently pre-processed data and can model several data blocks. However, one of the limitations of the PORTO is that it still involves a sequential step of first modelling common information and later the unique information from differently pre-processed data. Hence, to avoid the sequential nature of both algorithms and to give equal chance to each differently pre-processed data block, the method PROSAC (Figure 1(c)) was proposed. In PROSAC, the ensemble of pre-processings is learned as a competition between information extracted from differently pre-processed data. At each step of PROSAC, one latent variable is extracted from differently pre-processed data blocks, and the pre-processing block minimizing the response residual is declared as the winner for that step. Note that just like SPORT, the PROSAC is also data scale independent because it also models each data block individually. PROSAC has the innate capability of modelling many data blocks due to its parallel nature of modelling individual data blocks. However, PROSAC also has a limitation in that it is a greedy approach which may sometime get stuck in the local minima during the model optimization. In general, like any chemometric method, careful optimization of SPORT, PORTO and PROSAC can allow achieving optimal models.

A comparison of SPORT, PORTO and PROSAC

In a recent study, a comparison of the three recent pre-processing ensemble methods was provided (Figure 2). The aim was to predict soluble solids content (SSC) in

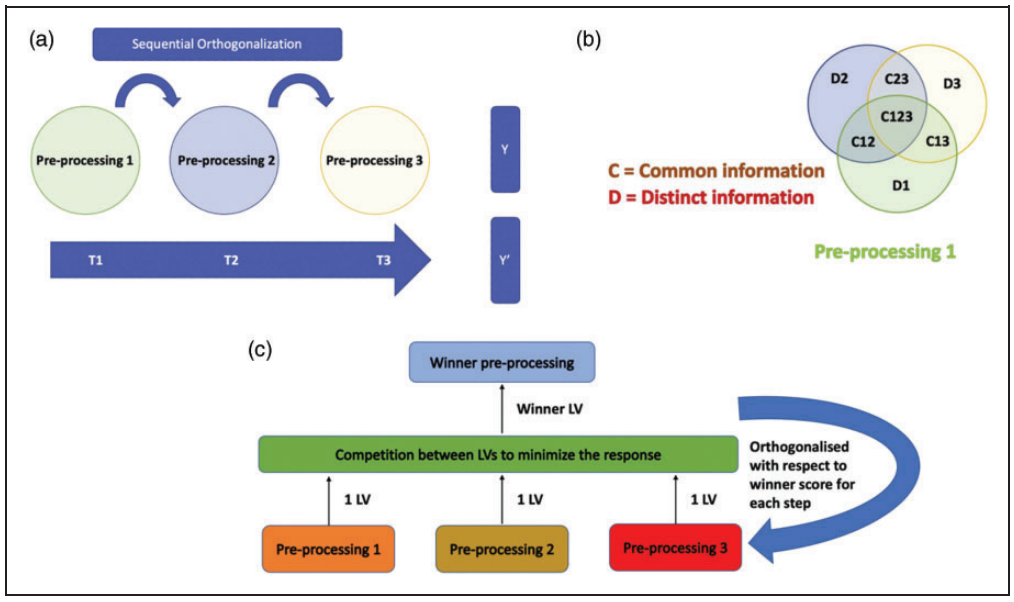

Figure 1. Schematic of different pre-processing ensemble approaches. (a) Sequential pre-processing through orthogonalization (SPORT), (b) parallel pre-processing through orthogonalization (PORTO) and (c) preprocessing ensemble with response oriented sequential alternation (PROSAC).

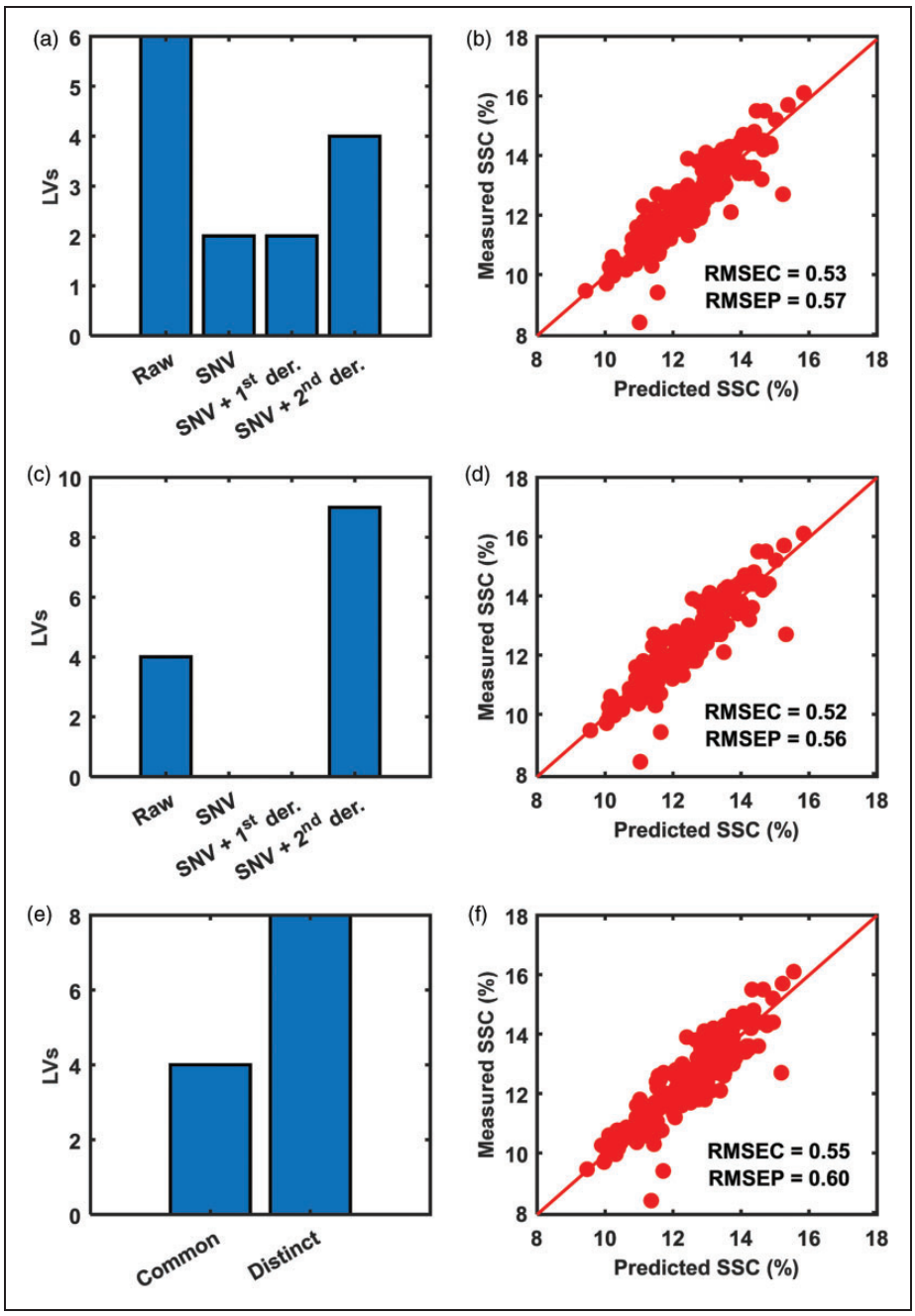

Figure 2. Performance of PROSAC, SPORT and PORTO on Pear data set.[4] (a) Winning block components for PROSAC, (b) prediction plot for PROSAC, (c) components from each block SPORT, (d) prediction plot for SPORT, (e) common and distinct components for PORTO and (f) prediction plot for PORTO.

fresh fruits using a handheld spectrometer. The spectral data were pre-processed with different pre-processings (Figure 2(a)), thereby making a total of four data blocks, one for raw data and three for different pre-processings such as normalization and derivatives. The predictive performances of three ensemble modelling approaches were comparable (Figure 2), as the prediction errors were close. One should note that in the traditional way of pre-processing optimization, the user must have explored models for individual pre-processing independently; however, with pre-processing ensemble, only one model was optimized. In Figure 2, one could also note how information was modelled from different blocks, particularly in the case of PROSAC and SPORT, where the model components for each pre-processing can be interpreted. For PORTO, the interpretation of model components is slightly different than the PROSAC and SPORT but can be performed if required. One key thing to note in SPORT and PROSAC analysis is that the model learned information from both the raw reflectance and differently pre-processed data. This is usually not possible when the aim is to select only a single best pre-processing approach. Often the practitioners neglect the useful information present in the raw data and directly aim to find the best pre-processing. In many

cases, the raw information contains useful patterns which contribute positively to explaining responses. For example, in fruit analysis, the parameter SSC is related to the ripeness level of the fruit. Along the ripening, both the physical (cellular structure) and chemical properties (macromolecules) of the fruit change; hence, one can assume that both scatter and absorption information are of use to explain the ripeness level of the fruit which is estimated as the SSC for many fruits. Another thing to note is that ensemble approaches such as PROSAC and SPORT also allow deselecting some pre-processings if they do not carry any complementary information compared to other pre-processings. For example, in the SPORT analysis (Figure 2(c)), two out of three pre-processings were deselected, as no model components were used from those pre-processings in the final model. In a practical scenario, this indicates that the user only needs to do one pre-processing of data and use it in an ensemble sequential model with raw data.

Conclusions

The selection of preprocessing has long been a challenge in NIR data modelling and that is why practitioners always aim to optimize and select the best pre-processing. However, preprocessing selection only focuses on selecting the preprocessing rather than exploring the complementary information present in differently preprocessed data for synergistic modelling. The new multiblock-inspired pre-processing ensemble methods allows one to learn complementary information that is usually expressed/refined by pre-processing with different techniques. One of the other benefits of using the ensemble pre-processing approaches for NIR practitioners is time saving that is usually required for exploring all pre-processings and their combinations independently. SPORT, PORTO and PROSAC, can be used interchangeably depending on the need and the knowledge of the data. For example, SPORT is highly suitable when a small number of pre-processings need to be explored and some knowledge on the order of their exploration is available. The PORTO technique can be used when no information on block order is available, and the aim is to use information from all data blocks to learn common and distinctive information. The PROSAC can be used when the aim is to learn an ensemble of several pre-processings while giving equal importance to all pre-processings. In terms of predictive performance, all methods are comparable, but some methods are more interpretable and insightful than others. For example, since SPORT and PROSAC aim to model data blocks individually, the user has much-refined access to the individual contribution of different pre-processings. Several of the pre-processing ensemble methods can be implemented with the free codes of Swiss-Knife PLS.[6]

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Mishra P, Biancolillo A, Roger JM, et al. New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends Anal Chem* 2020; 132: 116045.
2. Mishra P, Roger JM, Jouan-Rimbaud-Bouveresse D, et al. Recent trends in multi-block data analysis in chemometrics for multi-source data integration. *TrAC Trends Anal Chem* 2021; 137: 116206.
3. Roger J-M, Alessandra B and Federico M. Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. *Chemometr Intell Lab Syst* 2020; 199: 103975.
4. Mishra P, Roger JM, Marini F, et al. Parallel preprocessing through orthogonalization (PORTO) and its application to near-infrared spectroscopy. *Chemometr Intell Lab Syst* 2021; 212: 104190.
5. Mishra P, Roger JM, Marini F. Pre-processing ensembles with response oriented sequential alternation calibration (PROSAC): a step towards ending the pre-processing search and optimization quest for near-infrared spectral modelling. *Chemometr Intell Lab Syst* 2022; 222; 104497.
6. Mishra P and Liland KH. Swiss knife partial least squares (SKPLS): one tool for modelling single block, multiblock, multiway, multiway multiblock including multi-responses and meta information under the ROSA framework. *Anal Chim Acta* 2022; 1206: 339786.