



Incremental and decremental active learning for optimized self-adaptive calibration in viscose production



Carlos Cernuda ^{a,*}, Edwin Lughofer ^{a,*}, Georg Mayr ^b, Thomas Röder ^c, Peter Hintenaus ^d, Wolfgang Märzinger ^e, Jürgen Kasberger ^f

^a Department of Knowledge-Based Mathematical Systems, Johannes Kepler University Linz, Austria

^b Kompetenzzentrum Holz GmbH, St.-Peter-Str. 25, 4021 Linz, Austria

^c Lenzing AG, 4860 Lenzing, Austria

^d Software Research Center, Paris Lodron University, Salzburg, Austria

^e i-RED Infrarot Systeme GmbH, Linz, Austria

^f Recendt GmbH, Linz, Austria

ARTICLE INFO

Article history:

Received 10 January 2014

Received in revised form 12 June 2014

Accepted 9 July 2014

Available online 17 July 2014

Keywords:

Viscose production

Near-infrared spectroscopy

Dynamic processes

On-line self-calibration

Decremental and incremental active learning

Cost reduction

ABSTRACT

In viscose production, it is important to monitor three process parameters as part of the spin bath in order to assure a high quality of the final product: the concentrations of H_2SO_4 , Na_2SO_4 and $ZnSO_4$. NIR-spectroscopy is a fast analytical method applicable to conditions of industrial production and is capable of determining those concentrations. The collective composition of the spin bath varies in the industrial process, which implies changes in the matrix of the aforementioned analytes. Thus, conventional static chemometric models, which are trained based on collected calibration spectra from Fourier transform near infrared (FT-NIR) measurements, show a quite imprecise behavior when predicting the concentrations of new on-line data. In this paper, we are presenting a methodology which is able to cope with on-line self-calibration and -adaptation demands in order to compensate high system dynamics, reflected in conceptual changes in the mappings between NIR spectra and target concentrations. The methodology includes intelligent strategies for actively selecting those samples which should be accumulated into and excluded from the current data window in order to optimize the generalization performance of calibration models (thus termed as *incremental and decremental active learning stages*) while keeping the number of update cycles (and thus required target measurements) as low as possible. This follows the company requirements in terms of necessary *cost reduction*. Experiments on real-world data streams from viscose production process show that the new self-calibration methods are able to significantly reduce the number of update cycles while still keeping the predictive quality of the calibration models high (below 5% errors) for H_2SO_4 and Na_2SO_4 . Incremental active learning is able to smoothen and improve the overall quality of the predictions, while decremental active learning achieves a lower number of medium to large prediction errors.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Motivation and problem statement

The viscose process is of economic significance and production has been growing rapidly for the past two decades. In the viscose production there is a demand for measuring the most important process parameters, namely the concentrations of H_2SO_4 , Na_2SO_4 and $ZnSO_4$. The acid and the two salts govern the precipitation and agglomeration of the cellulose from viscose solution and the formation of the viscose

fiber. The concentration of those components has major influence on the fiber properties. Hence, the accurate knowledge and control of the concentration are a prerequisite for the production of high-quality viscose fibers in the industrial processes. The conventional method to determine those concentrations is titration [1], which is quite time consuming, although it can be implemented in an automated fashion, i.e. automated titration draws a liquid sample from the process line and accomplishes an analysis on a regular basis. This is the case in the process under consideration in this work. The automated titration delivers an analytical value each 8th minute, whereas NIR-spectroscopy is capable to deliver a spectrum and consequently an analytical result each 10th second. The spin bath in the spinning trough is recirculated approximately each 20 min, compared to which the analysis time of the conventional titromat is quite long, what results in long reaction times to adjust the spin bath composition. This reaction time for adjustment could be dramatically shortened by the fast analysis achievable by

* Corresponding authors.

E-mail addresses: carlos.cernuda@jku.at (C. Cernuda), edwin.lughofer@jku.at (E. Lughofer), geo.mayr@lenzing.com (G. Mayr), t.roeder@lenzing.com (T. Röder), ph@i-red.at (P. Hintenaus), wm@i-red.at (W. Märzinger), juergen.kasberger@recendt.at (J. Kasberger).

NIR-spectroscopy. With the usage of chemometric models based on NIR spectra, a more frequent supervision of the process is possible. Furthermore, the automated titration also entails considerable costs for chemicals and maintenance, thus the reduction of its usage by automatic quantification based on chemometric models is highly desired. A complete substitution of the automated titration would be the ideal situation, however due to the high system dynamics of the spin bath, self-adaptation of the chemometric models is expected to be required from time to time (which could be verified in a former publication [2]). This is mainly because operating modes and states of the industrial processes might vary according to different products (i.e. fiber types) or raw materials processed, hence also the overall composition of the spin bath in detail might vary. E.g. the spin bath might contain additives or degradation products of the cellulose to a varying content [3], what is also conditioned by the dynamics of the industrial continuous operation mode and the recirculation of process media. These fluctuations in spin bath composition represent changes of the matrix of the analytes H_2SO_4 , Na_2SO_4 and ZnSO_4 . For NIR-spectroscopy these fluctuations might represent a problem, as a NIR-spectrum contains comprehensive information about the chemical composition, and specific analytical information is extracted by chemometric models. Due to the high expense for performing titrations, a requirement of the company is to reduce measurements and thus the update cycles. Ideally, the goal would be only to measure and update the model 3–4 times a day, i.e. each 6th to 8th hour instead of each 8th minute. Summarizing, we should achieve a double requirement consisting of capturing the high dynamics of the system, as well as reducing costs by decreasing the frequency of the titrations.

1.2. State-of-the-art and open problems

Ideally, when it comes to cost reduction, the best option would be to apply batch models. These are usually trained once based on off-line calibration samples, which are supposed to be representative of the whole process under study, and then applied to new on-line data for quantification and prediction of substances. Unfortunately, the dynamics of these matrix effects abandons the application of standard chemometric modeling tools, as e.g. can be found in [4,5] or [6] and have been applied to viscose production in the past [7,8]. In fact, in [2], it could be verified that, when using conventional modeling methods such as PLS [9], iPLS [10], PCR [11], LWR [12] and others (most of them used as implementations in the PLS-Toolbox¹) in the viscose production process under our study, the model errors start to significantly drift over time. For the first few hundred of new on-line samples, the performance was similar to the one obtained on off-line training samples, but the more deteriorated the more time passed by (\rightarrow drift effect). Thus, focusing on capturing the dynamics of the process, in [2] a new concept was developed, termed as *evolving chemometric models* (*eChemo*) which possess the ability to automatically self-adapt and re-calibrate based on newly recorded on-line NIR spectra samples with minimal resources. The main characteristics of *eChemo* were:

- It required an initial model (a non-linear Takagi-Sugeno fuzzy system [13]) from some pre-recorded and off-line stored calibration samples. The parameters of the initial models (especially structure, i.e. number of rules, and input dimensionality) were optimized using a best-parameter grid search scenario coupled with cross-validation and specific model selection scheme, see [14].
- It was able to adapt the final selected model with new incoming on-line samples in a single-pass incremental manner, based on target concentrations measured by the automated titration. An interesting finding was that the model updates require evolution (and back-

pruning) of structural components on demand [15] as well as forgetting of older samples [16,17] in order to increase both, compactness [18] and flexibility (condemning the usage of recursive linear models as used before in [19] and [20]). This assured that the model was always freshly up-to-date during the on-line process and a fast learning could be achieved.

- It performed an update with each new incoming sample (approx. each 8th minute), thus requiring permanent values from the automated titration. The quality of the predictions remained on a very high level (max. 3% relative error) during the whole process, thus prevented any drift effects. Regarding cost reduction, as can be seen in [2], when testing *eChemo* on updating each 8th or each 16th sample (far even from the company requirements of adapting 3–4 times a day, i.e. each 60th–45th sample) the performance loss is huge.

Therefore, neither the State-of-the-art batch models nor *eChemo* could find an acceptable trade-off between accuracy and cost. The former because of low accuracy due to the drift effect, and the latter because of an unacceptably high cost within a dynamically changing environment. Thus, new approaches should be investigated.

1.3. Our approach

In this paper, we demonstrate a different approach which is based on self-adaptive calibration models within a sliding window concept (Section 3.1), which puts more flexibility in the incremental learning phase and especially also in the forgetting process of older samples. We include more control on the issue which samples to forget than in the usual forgetting strategies, only based on the age concept (timeline). This is achieved by a *decremental active learning stage* (Section 4.2) which selects that sample from the window with lowest information gain in terms of diversity and model quality. Furthermore, the sliding window concept allows the introduction of more flexibility in terms of self-adjusting learning parameters and self-optimizing input dimensionality for the self-adaptive calibration models, thus allowing the integration of an input structure change, which is not possible in batch modeling nor in *eChemo* (in which the input structure is optimized in an initial off-line calibration phase, and then is kept fixed for the on-line phase). Finally, in order to put control on the samples requested for the self-adaptive calibration models and to avoid unnecessary updates not bringing any performance (e.g. samples already included in the model updates), we pursue a strategy which we term as *incremental active learning stage*. Therefore, we do not use each equidistant sample to be included in the sliding window, but we actively select samples with the highest information gain compared to the samples already present in the current window, i.e. samples which are expected to be able to generalize the chemometric models and thus to improve their predictive quality on new query points most. The number of actively selected samples has to be constrained in order to meet the requirements by the company, namely to record maximally 3 to 4 target values per day (thus allowing maximal 3 to 4 updates of the sliding window per day). As model architectures, we will exploit conventional PLS method (from the PLS-toolbox) and fuzzy systems with the batch model variant FLEXFIS as learning engine (as used in [2]), but employing latent variables from PLS as inputs. Thus, it is termed as FLEXFIS + PLS which can be seen as a sort of non-linear version of PLS; this will provide us with the information whether non-linearity in the self-adaptive calibration models helps to improve the performance.

The results include a comparison of the new sliding window approach with the original *eChemo* models [2], with different variants for deleting samples from the current windows (see Section 6), namely oldest versus random versus our enhanced selection strategy within the scope of *decremental active learning* as well as with different variants for accumulating samples in the window, namely blind static equidistant (default sliding window approach) versus the adaptation of an approach in [21] (based on Euclidean distances) versus our enhanced

¹ http://www.eigenvector.com/software/pls_toolbox.htm.

Download English Version:

<https://daneshyari.com/en/article/1180729>

Download Persian Version:

<https://daneshyari.com/article/1180729>

[Daneshyari.com](https://daneshyari.com)