ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Utilization of Chemical Structure Information for Analysis of Spectra Composites

Kristin Domaschke[1,2], André Rossberg[1] and Thomas Villmann[3] *

1- Helmholtz-Zentrum Dresden-Rossendorf, Institute of Resource Ecology, P.O. Box 51 01, 01314 Dresden, Germany

2- current address: Life Science Inkubator Sachsen GmbH & Co. KG, project team NanoscopiX, Tatzberg 47, 01307 Dresden, Germany

3- University of Appl. Science Mittweida - Dept. of Mathematics, Technikumplatz 17, 09648 Mittweida, Germany

Abstract. In this paper, we propose the utilization of structural information of spectral data during the preprocessing to extend the ability of subsequent analysis methods. Specifically, we expect a dataset of measured spectra containing mixtures of only a few spectral components. Using the concentration ratios for a small subset of mixtures and the chemical structural knowledge, theoretical spectral components are generated. Then a set, which combines measured and theoretical spectra, is analyzed using a self-organizing map to predict the unknown mixture ratios of the remaining subset by an associative learning. At this time, the initial study on simulated data reached very good results.

1 Introduction

The prediction of concentration of mixture components is an important task in chemical data analysis. In this paper we propose an approach based on Kohonen's Self-Organizing Map (SOM,[1]) and an utilization of structural chemical knowledge, if the mixture information is available for a few data of the whole data set.
The SOM is one of the most popular data mining and visualization methods for vectorial data. It provides a non-linear mapping of the possibly high-dimensional data onto a low-dimensional, frequently two-dimensional, grid. Under certain conditions this mapping preserves the topological properties of the data [2].
Chemical spectral data are known to be high-dimensional in general and, therefore, are predestined for processing using SOMs in order to visualize them or to detect clusters. To analyze the dependence of the spectral structure of data on specific chemical parameters, such as the pH-values or concentrations, a fused mapping of the spectral data together with further information is frequently requested. For this need, Melssen proposed the XY-fused SOM [3]. In this approach, two SOMs, one for spectral data and one for chemical parameters, are combined bi-directionally via a merged winner determination. Yet, this structure is very complex and fails if only a small amount of data is available for learning, as it is frequently the case in chemical data analysis.

*The authors would like to thank Dr. Vinzenz Brendler and Dr. Kay Grossmann, of Institute of Resource Ecology at Helmholtz-Zentrum Dresden-Rossendorf, for hosting this research.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Here, the factor analysis, which is used as standard method for spectral decomposition, can be mentioned. Nevertheless, the non-linearity of SOMs seems to have much greater potential especially for further research on spectral unmixing problems as the factor analysis.
Thus, in this paper we propose the following strategy: First, in a preprocessing step, we use the available mixture information for a small subset of the spectral data to estimate theoretical pure component spectra, which can serve as additional data vectors with the identity matrix as known mixture information. For this purpose, chemical structure information is explicitly taken into account. Thereafter, we apply associative learning in self-organizing maps to estimate the unknown mixture ratios for all spectra of the data set. In the last section of this paper we present the application of this strategy to a simulated data set.

2 Preprocessing of spectra using chemical structure information

2.1 Spectral pattern and structural information in composite spectra

We assume n spectra $\mathbf{d}_i$ obtained from some measurements with N_B bands collected in the data matrix $\mathbf{D} = \{\mathbf{d}_1, \cdots, \mathbf{d}_n\} \in R^{n \times N_B}$. In our data to be analyzed, these spectra are obtained by extended X-RAY absorption fine structure spectroscopy (EXAFS,[4]) as *absorption spectra*. These measured spectral composites $\mathbf{d}_i$ consist of sparse mixtures of m theoretical components, where $m << n$. These m pure spectra, which are estimated as described later, are collected in the matrix $\mathbf{R} \in R^{m \times N_B}$ with $\mathbf{R} = \{\mathbf{r}_1, \cdots, \mathbf{r}_m\}$ and $\mathbf{r}_j \in R^{N_B}$. Further we denote by $\mathbf{C} \in R^{n \times m}$ with $\mathbf{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_n\}$ and $\mathbf{c}_i \in R^m$ the unknown component concentration ratios of the component spectra $\mathbf{R}$ within the composites $\mathbf{D}$.
The structural information of the spectral signal can be described by the Lambert-Beer law [5]. In the following, this definition is explained for absorption spectroscopic methods, but can be adapted simply to many other spectroscopic methods, such as fluorescence spectroscopy.
The spectral signal of an absorbing substance is called extinction E in absorption spectroscopy and depends proportionally on the thickness of the sample b, the concentration c and on an extinction coefficient α by

$$E(\lambda_k) \quad = \quad b \cdot c \cdot \alpha(\lambda_k) \quad \left([\mathrm{cm}] \left[\frac{\mathrm{mol}}{\mathrm{l}} \right] \left[\frac{\mathrm{l}}{\mathrm{mol} \cdot \mathrm{cm}} \right] \right). \tag{1}$$

The unit of the extinction coefficient α results from the units for b and c. We suppose the thickness of the sample b is standardized to 1cm for each measurement, so we can reduce the equation (1) to

$$E(\lambda_k) = c \cdot \alpha(\lambda_k) \quad \left(\left[\frac{\mathrm{mol}}{\mathrm{l}} \right] \left[\frac{\mathrm{l}}{\mathrm{mol}} \right] \right)$$

If we further assume a high concentration of $1 \frac{mol}{l}$ for the substance, we see that the extinction coefficient $\alpha(\lambda_k)$ is equal to $E(\lambda_k)$ and therefor has the same data characteristics as $E(\lambda_k)$.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Furthermore, mixture samples contain more than one absorbing substance, which here are assumed having no interactions between them. Then, the absorption of a composite sample E_i, with i is the index of measurements in $1 < i < n$, is the sum of the singular extinctions

$$E_i(\lambda_k) = E_1(\lambda_k) + E_2(\lambda_k) + \cdots + E_m(\lambda_k)$$

resulting in the linear Lambert-Beer law for the m components.

$$E_i(\lambda_k) = \sum_{j=1}^{m} c_{ij} \cdot \alpha_j(\lambda_k) \tag{2}$$

Furthermore, these equations show, that the extinction E and their extinction coefficient α relies on λ_k , which is the wavelength of the k_{th} band with $k = \{1, \cdots, N_B\}$. Depending on the dimension of λ_k, which can characterize either only one or all N_B spectral bands, we get either a single extinction value E or an extinction vector $\mathbf{E}$. This also applies to the extinction coefficients α.
For our data set, the extinction vector $\mathbf{E}$ of a sample i is the measured spectrum $\mathbf{d}_i$. Further, the m pure components can be identified with their extinction coefficients $\boldsymbol{\alpha}_j$ such that $\mathbf{r}_j = \boldsymbol{\alpha}_j$ is valid. Thus we obtain in (2)

$$\mathbf{d}_i(\lambda_k) = \sum_{j=1}^{m} c_{ij} \cdot \mathbf{r}_j(\lambda_k) \tag{3}$$

$$\mathbf{D} = \mathbf{CR}\,. \tag{4}$$

We observe that under certain conditions the measured spectra $\mathbf{d}_i$ can be determined additively by the spectra of the pure components $\mathbf{r}_j$ and their concentration values c_{ij} within the composites. For a set of measured spectra $\mathbf{D}$ equation (3) holds in the same way, as shown in (4). Further we only consider convex data structures, i.e. $\sum_{j=1}^{m} c_{ij} = 1$. In particular, it is assumed that the number of all contained components is known in advance.
Unfortunately, the concentration matrix $\mathbf{C}$ is generally not known for the measured data. Hence, we cannot calculate the matrix $\mathbf{R}$. However, the deduced structural relation (4) can be used as chemical structure information for structured preprocessing if a few concentrations are available.

2.2 Using structural information for structured preprocessing

In our problem we assume a spectral data subset $\mathbf{D}_S \in R^{m \times N_B}$, for which the concentrations $\mathbf{C}_S \in R^{m \times m}$ are known. Thus, if we have this additional data at least for as many spectra as components $\mathbf{r}_j$ with $1 < j < m$ underlying the data, we can compute the pure component spectra $\mathbf{R}$ using the structural information in (4). For this purpose, we apply an iterative Jacobi scheme [6], such that for each iteration step $t + 1$ we obtain

$$\mathbf{R}^{t+1} = \mathbf{O}^{-1}\left[\mathbf{D}_S - \mathbf{L}\mathbf{R}^t\right] \quad (t = 1, 2, \cdots)$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

where the following notation is used.

$$\mathbf{C}_S = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{pmatrix} = \mathbf{L} + \mathbf{O} \text{ with}$$

$$\mathbf{L} = \begin{pmatrix} 0 & c_{12} & \cdots & c_{1m} \\ c_{21} & 0 & \cdots & c_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & 0 \end{pmatrix}, \quad \mathbf{O} = \begin{pmatrix} c_{11} & 0 & \cdots & 0 \\ 0 & c_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_{mm} \end{pmatrix}$$

The iteration, starting with some set values $\mathbf{R}^0$, is performed until $\|\mathbf{R}^{t+1} - \mathbf{R}^t\| < \epsilon$, where $\epsilon > 0$ is a chosen tolerance, is reached. Here, $\|\cdot\|$ denotes the spectral norm. Convergence is guaranteed if either

$$\max_i \sum_{j=1, j\neq i}^{m} \left| \frac{c_{ij}}{c_{ii}} \right| < 1 \text{ or } \max_j \sum_{i=1, i\neq j}^{m} \left| \frac{c_{ij}}{c_{ii}} \right| < 1$$

holds [6]. As a result, we obtain the estimates $\mathbf{R}$ of the pure component spectra. The concentration matrix $\mathbf{C_R}$ with $\mathbf{C_R} = \{\mathbf{e}_1, \cdots, \mathbf{e}_m\}$ and $\mathbf{e}_j \in R^m$ is simply the unity matrix of dimension m consisting of the respective unity vectors $\mathbf{e}_j$. In this way, we are able to generate additional spectra with known concentrations based on the chemical structure information (4).

3 Utilization of the extended data set to SOM-association-learning

After the structured preprocessing, we have, on the one hand, the originally given data, consisting of $\mathbf{D}$ with the subset $\mathbf{D}_S$, for which the concentrations $\mathbf{C}_S$ are available. W.l.o.g. we suppose these spectra to be $\mathbf{d}_1 \ldots \mathbf{d}_m$ with concentration vectors $\mathbf{c}_1 \ldots \mathbf{c}_m$. On the other hand we generated the pure component spectra $\mathbf{R}$ with trivial concentrations $\mathbf{C_R}$. Thus we have the following dataset:

$$\mathbf{V} = \begin{pmatrix} d_{1,1} & \cdots & d_{1,m} & r_{1,1} & \cdots & r_{1,m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ d_{N_B,1} & \cdots & d_{N_B,m} & r_{N_B,1} & \cdots & r_{N_B,m} \\ c_{1,1} & \cdots & c_{1,m} & e_{1,1} & \cdots & e_{1,m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_{m,1} & \cdots & c_{m,m} & e_{m,1} & \cdots & e_{m,m} \end{pmatrix}, \tilde{\mathbf{V}} = \begin{pmatrix} d_{1,m+1} & \cdots & d_{1,n} \\ \vdots & \ddots & \vdots \\ d_{N_B,m+1} & \cdots & d_{N_B,n} \\ c_{1,m+1} & \cdots & c_{1,n} \\ \vdots & \ddots & \vdots \\ c_{m,m+1} & \cdots & c_{m,n} \end{pmatrix}$$

Hence, the *fused* dataset $\mathbf{V} = \{\mathbf{v}_1, \cdots, \mathbf{v}_{m+m}\}$ contains both informations, the spectra and their known concentrations merged into single vectors $\mathbf{v}_l \in R^p$ with $p = N_B + m$ and $1 < l < m+m$. The set $\tilde{\mathbf{V}}$ collects the remaining $n-m$ spectra together with the unknown concentrations $\mathbf{c}_{m+1} \ldots \mathbf{c}_n$.
Both datasets are used for associative learning in SOMs [1]. In our application, the SOM grid $\mathbf{A}$ is a square lattice of edge length N such that the coordinate vectors are $\mathbf{k} \in N^2$. The weight vectors or prototypes $\mathbf{W} = \{\mathbf{w_k}\}$ assigned to

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

$\mathbf{A}$, have the same dimension p as defined for the data vector of $\mathbf{V}$.
Usual SOM training takes place as applying the winner determination

$$\mathbf{s}(\mathbf{v}) = \arg\min_{\mathbf{k}\in A} \|\mathbf{v} - \mathbf{w_k}\|$$

and subsequent prototype adaptation according to

$$\Delta\mathbf{w_k} = \epsilon h_{\mathbf{ks}} (\mathbf{v} - \mathbf{w_k}) \quad \text{with} \quad h_{\mathbf{ks}} = \exp\left(-\frac{\|\mathbf{k}-\mathbf{s}\|_A}{2\sigma^2}\right)$$

being the neighborhood function. Here $0 < \epsilon \ll 1$ is the learning rate and σ is the neighborhood function. To keep the information of the subset $\tilde{\mathbf{V}}$ we also feed these vectors into the training scheme. Because of the unknown concentration information for these vectors $\tilde{\mathbf{v}}_k$ only the spectral information is used for winner determination. For prototype update we apply the spectral information and the concentration estimations given by the current best matching unit.
After training, the unknown concentrations for the vectors $\tilde{\mathbf{v}}_k$ can be estimated by association, i.e. the winner $\mathbf{s}(\tilde{\mathbf{v}}_k)$ is determined and we simply set $\tilde{v}_k^{N_B+i} = w_{\mathbf{s}(\tilde{\mathbf{v}}_k)}^{N_B+i}$ for $i = 1 \ldots m$ as association step.

4 Application of structured information processing for an EXAFS data set

For this application a dataset of $n = 14$ *simulated* EXAFS spectra with $N_B = 100$ is used. The underlying chemical structure information assumes $m = 4$ components and a sinusoidal pure component spectrum according to $\sin(b_j \cdot \mathbf{f})$ for a given frequency vector $\mathbf{f} \in R^{N_B}$ [7]. Here we consider $\mathbf{b} = (0.125, 0.25, 0.5, 1)$ for the different pure component spectra. We generated the simulated spectra as mixtures of these pure components with a predefined mixing matrix $\hat{\mathbf{C}}$ of concentrations.
In the simulations we provided the concentrations only for $m = 4$ simulated spectra to the above described processing algorithm using a SOM-grid A with $N = 15$. Thus, these data play the role of the subset $\mathbf{D}_S$. For the SOM-training the neighborhood range σ remains non-vanishing also during the convergence phase of SOM-learning to keep the generalization ability of the model despite the small number of available training data, which is a common procedure for those cases [8]. Further, we applied a prototype correction to ensure the relation (4) for the prototypes during the learning process. Thus, the structural knowledge is used again.
After the application of the complete analysis procedure we obtain the estimated concentrations for the remaining 10 spectra. We can compare them with the original ones used for generating the simulated data set. As we can see in Fig. 1, we get accurate predictions. Thus, the association learning by SOMs together with the utilization of the chemical structure information is able to estimate the unknown concentrations quite well for this data set.

5 Conclusion

We propose a data analysis framework for spectral data processing in order to estimate mixture concentrations in spectra composites under the assumption of

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

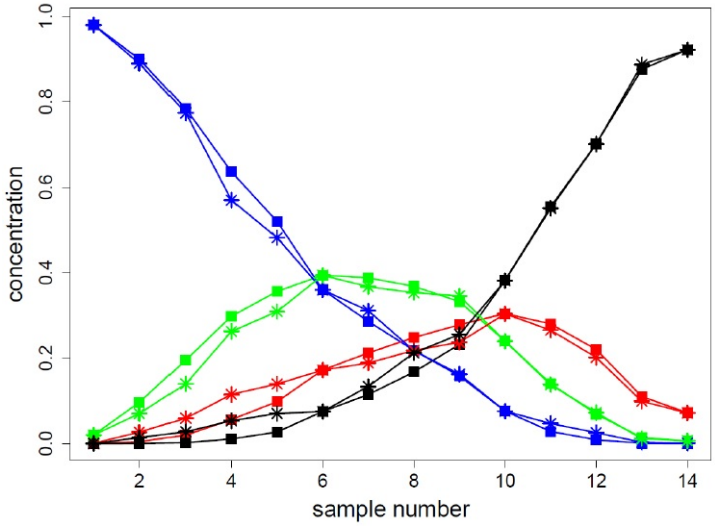

Fig. 1: Comparison of the 4 original component concentrations (■) and the respective predicted values (⋆) obtained from the structured data processing for the 10 spectra not contained in $\mathbf{D}_S$.

sparsely available concentration information. This approach takes the structural chemical information into account to generate an extended while reliable data set, which then can be used for association learning in SOMs. We introduced the method and explained, how the structural-chemical information is integrated. We exemplary demonstrated the method for an artificial dataset. Yet, further simulation and stability analysis is required. In particular, the influence of noisy data has to be considered. Thus, the presented work is only a proof of concepts so far, which has to be evaluated further in the next steps of our research.

References

[1] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
[2] Th. Villmann, R. Der, M. Herrmann, and Th. M. Martinetz. Topology preservation in self-organizing feature maps: Exact definition and measurement. *IEEE Transactions on Neural Networks*, 8(2):256–266, 1997.
[3] W. Melssen, R. Wehrens, and L. Buydens. Supervised Kohonen networks for classification problems. *Chemometrics and Intelligent Laboratory Systems*, 83(2):99–113, 2006.
[4] A. Rossberg, T. Reich, and G. Bernhard. Complexation of uranium (vi) with protocatechuic acid - application of iterative transformation factor analysis to exafs spectroscopy. *Analytical and bioanalytical chemistry*, 376(5):631–638, 2003.
[5] D.A. Skoog and J.J. Leary. *Instrumentelle Analytik: Grundlagen - Geräte - Anwendungen*. Springer-Lehrbuch. Springer, 1996.
[6] H. Friedrich and F. Pietschmann. *Numerische Methoden: ein Lehr- und Übungsbuch*. Walter de Gruyter GmbH & Co. KG, 2010.
[7] K. Domaschke. Multidimensionale Spektrenanalyse mittels neuronale Netze. diploma thesis, Hochschule Zittau/Görlitz - Univerity of Applied Science, 2012.
[8] Th Villmann. Neural maps for faithful data modelling in medicine - state of the art and exemplary applications. *Neurocomputing*, 48(1-4):229–250, 2002.