ISSN 1060-992X, Optical Memory and Neural Networks (Information Optics), 2015, Vol. 24, No. 2, pp. 93–101. © Allerton Press, Inc., 2015.

Comparison of the Quality of Solving the Inverse Problems of Spectroscopy of Multi-Component Solutions with Neural Network Methods and with the Method of Projection to Latent Structures

A. O. Efitorov[a, b], S. A. Burikov[a, b], T. A. Dolenko[a, b], I. G. Persiantsev[a], and S. A. Dolenko[a]

[a] *Skobeltsyn Institute of Nuclear Physics, Moscow State University, Russia*

[b] *Faculty of Physics, Moscow State University, Russia*

e-mail: sasha.efitorov@yandex.ru, dolenko@srd.sinp.msu.ru

Received in final form, February 24, 2015

Abstract—This study provides comparative analysis of application of artificial neural networks and method of projection to latent structures (partial least squares) for simultaneous determination of types and concentrations of dissolved inorganic salts in multicomponent water solutions by Raman spectra. It is shown that the method of projection to latent structures has several advantages, such as the quality of the solution and the time of construction of a regression model, when solving problems with low level of nonlinearity.

Keywords: neural networks, projection to latent structures, partial least squares, inverse problems, spectroscopy, identification

DOI: 10.3103/S1060992X15020022

INTRODUCTION

It is well known that artificial neural networks (NN) are a class of mathematical algorithms that demonstrate high efficiency in solving problems of approximation, forecasting, evaluation, classification and pattern recognition. NN are also widely used in solving inverse problems (IP), where their properties such as training by example, high noise immunity, resistance to contradictory data [1, 2] play a special role. In addition to NN, the methods that are able to solve problems of this type are projection methods. One of the most efficient of them is the method of projection to latent structures (PLS), or partial least squares [3], which has been successfully used to build regression models and classifiers [4–7]. This paper presents comparison of methods for IP solving at the example of the complex IP of determination of types and partial concentrations of inorganic salts in multicomponent aqueous solutions by their Raman spectra.

The problem of determining the concentrations of substances dissolved in water is very important for oceanography, environmental monitoring and control of mineral, industrial and waste waters. This problem requires to be solved in a non-contact express mode with acceptable accuracy.

Raman spectroscopy method satisfies these requirements. The principle possibility of using Raman spectra for diagnostics of solutions is due to their high sensitivity to the types and concentrations of salts dissolved in water. In [8, 9], it is proposed to use the Raman spectra of complex ions (such as bands of NO^{3-}, SO_4^{2-}, PO_4^{3-}, CO_3^{2-} anions near 1000 cm^{-1}) to determine the types and concentrations of salts in water. Anion type can be determined by the position of its spectral band, its concentration—by its intensity. However, this method can only be used for analysis of substances with their proper Raman bands, i.e. of salts with complex ions. The authors of [10–13] have suggested and developed methods for determining concentrations of dissolved salts by Raman valence band of water, including methods based on artificial neural networks [12, 13].

The method of identification and determination of individual concentrations of salts, used in this paper, can work with salts containing both complex and simple ions. The method was first proposed by the authors in [14] and developed in [15–18]. Presence of complex ions in the solution is most easily deter-

[1] This study has been supported by the Russian Science Foundation grant no. 14-11-00579.

mined by the presence of their peaks in the low-frequency region of the Raman spectrum, and their concentrations are determined by the dependence of intensity of these bands on the concentration, taking into account the influence of other salts on this intensity. Recognition and determination of concentrations of simple ions is performed by the change of shape and position of water Raman valence band in presence of all salts dissolved in water.

Simultaneous determination of individual concentrations of a number of dissolved ions and their identification are provided by the use of a NN, which performs simultaneous analysis of both areas of the Raman spectrum (valence and low frequency ones). Application of formal mathematical models is expedient due to the fact that because of complexity of the object there is no adequate physical model capable of numerical modeling of the dependence of Raman spectrum of the solution on the concentrations of the dissolved salts, especially taking into account their nonlinear interaction. Therefore, in this paper we apply the "experiment-based" methodical approach to the solution of the inverse problem [19]. Within this approach, the data used to build formal models is obtained experimentally, requiring a sufficiently large number of measurements. To implement this approach, the authors have obtained 8695 experimental spectra for 4268 different solutions.

Within Method 1, a single NN has as many outputs as there are salts that are determined, and the amplitude at these outputs is proportional to the concentrations of the corresponding salts. In [15], a method for solving the IP using NN in two stages (Method 2) was described. At the first stage, a single common NN detects the component composition of the solution, at the second stage another specialized NN determines the concentrations of the detected components. This approach was also used in the present study. In [15] it has been demonstrated that with non-aggregated data (see below) the detection of the salts at the first stage was nearly always correct, but the results of determination of the concentrations of the components was in general worse than within Method 1. Apparently, this is due to the extremely unfavorable ratio of the number of samples and the number of input features for a separate class corresponding to a particular component composition of the solution.

DATA PREPARATION

A diagram and a description of the experimental setup are provided elsewhere [14, 15].

The objects of research were aqueous solutions of salts with significant content in natural waters—NaCl, NH_4Br, Li_2SO_4, KNO_3, CsI. The concentration of each salt in the solution varied in the range from 0 to 2.5 M with concentration step of 0.2–0.25 M.

Initially, each band of the Raman spectrum was recorded into the range 1024 spectral channels wide, in the frequency range 200–2300 cm^{-1} for the low-frequency band, and 2300–4000 cm^{-1} for the valence band. For further processing, more narrow informative ranges were selected: 766 channels in the range 281–1831 cm^{-1} for the low frequency band, and 769 channels in the range 2700–3900 cm^{-1} for the valence band.

Next, the horizontal pedestal caused by light scattering in the cuvette with the sample was subtracted separately from each band, and each band was normalized to the area of the valence band in the specified informative range. Then, linear aggregation (summation and averaging) of intensity values in each eight neighboring spectral channels has been performed over the entire spectral range. In [18], it has been demonstrated that aggregation was the most efficient method of input data dimensionality reduction, and that it improved the quality of IP solution.

The resulting data array (192 features, 9144 samples) was divided into 31 "combinatorial" classes, differing from each other by the composition of the salts present in the solution. This array was used to solve the problem of determining the concentrations of the dissolved salts by NN Method 2 and by PLS method. In its turn, the array not divided into classes was used to determine the components present in the solution at the first stage of NN Method 2, as well as to determine the types and concentrations of components within NN Method 1.

Note that PLS is a linear method. That's why one of the variants of data preprocessing was as follows: $10^{(-x)}$, where x was the value of intensity in each spectral channel. This operation transformed transmission spectra to absorption spectra. However, using the transformed data with neural network methods didn't improve the results of the IP solution, because the neural network is a nonlinear approximator itself.

Further processing was separation of the data within each combinatorial class randomly into training, test and examination sets in the ratio of 70 : 20 : 10, respectively. Sets of each type for all classes were also united to obtain the set of this type for the full data array. The training set was used to build regression models. The test set was used to prevent NN overtraining – when the mean squared error on the test set begins

to increase, NN training should be stopped. The formation of the PLS-regression model was stopped when convergence on the training set was achieved. To provide an independent estimation, all the results presented below are for the examination set.

STATEMENT OF THE PROBLEM

Based on the presented results and assumptions [15], we solved the IP on aggregated data with two neural network methods described above (in one or in two stages). The assumption on the efficiency of the second stage of Method 2 (determination of concentrations) is based on the increase in the ratio of the number of samples to the number of features for an order of magnitude after aggregation. Also, the determination of the concentrations of components for each corresponding combinatorial class was conducted by PLS.

In all the computational experiments described below, the neural network solution of the problem was obtained with the perceptron with three hidden layers containing 40, 20 and 10 neurons. It should be noted that other computational experiments using perceptrons with different numbers of hidden layers and neurons were also conducted, but the architecture 40–20–10 has shown the best results. In the output layer, linear activation function was used, in the hidden layers—logistic activation functions. The following training parameters were used: learning rate 0.01; learning moment 0.5; stopping criterion—1000 training epochs after minimum error on the test data set. To eliminate the influence of the NN weights initialization on the result, 5 neural networks with various initial approximations of weights were trained in each case, and their results of the IP solution were averaged.

PLS model was based on the NIPALS algorithm for 25 components. The stopping criterion was the convergence of the algorithm on the training set. During IP solution, the smallest value of the mean absolute error (MAE) was chosen for the given salt in the given combinatorial class depending on the amount of the components used.

COMPUTATIONAL EXPERIMENTS AND THEIR RESULTS

Table 1 presents the values of MAE for determination of the concentrations of each salt for each combinatorial class obtained with NN Method 1; Table 2—those obtained with NN Method 2; Table 3—those obtained by PLS method on initial data, Table 4—those obtained by PLS on data preprocessed by transformation of transmission spectra to absorption spectra. The cells marked by filling of two corners contain the smallest value of the MAE among the results of use of all methods; those marked with diagonal stroke contain the best result of the solution obtained by more than one method.

In general, the worst IP solution result was demonstrated by NN Method 2, where a separate NN was trained for each combinatorial class. The value of the average (over all salts and classes) total MAE was 0.02787; the best results of determining salt concentrations in comparison with other methods were observed in only 14 of 80 cases, the absolute best in 11 cases (see Table 2).

More convincing results of the IP solution were achieved by the universal neural network trained simultaneously to identify the composition of the solution and to determine the concentrations of the components. The value of the average total MAE was 0.02605, the best results of determining salt concentrations in comparison with other methods were observed in 29 of 80 cases, the absolute best in 23 cases (see Table 1).

The method of constructing the PLS-regression models for each combinatorial class on initial data was somewhat more efficient. The value of the average total MAE was 0.02453; the best results were demonstrated in 32 of 80 cases, the absolute best in 11 cases (Table 3).

The best results were demonstrated by the method of constructing the PLS-regression models for each combinatorial class using preprocessed data. The value of the average total MAE was 0.02378; the best results were demonstrated in 34 of 80 cases, the absolute best in 10 cases (Table 4).

It should be noted that if there were 5 dissolved salts, the PLS-regression method has shown worse IP solution results in comparison with the results of application of neural network methods, both of which in this case showed similar results. This can be explained by the fact that PLS-regression is a linear regression method, thus it gives a poor description of strong nonlinear interactions among the components of the solution. For the same reason, in the case of a single-component solution, the quality of the IP solution by PLS is much higher than the quality of solution of the same problem obtained by a neural network. A nonlinear preprocessing of data allowed improving the results of application of PLS-regression method to complex water solutions.

Table 1. The mean absolute error (MAE) on examination set for determination of the concentrations of salts in different combinatorial classes by neural network Method 1. The marked results are best among all the methods

Class number	Number of salts	NaCl	NH_4Br	Li_2SO_4	KNO_3	CsI
1	1	0.054				
2	1		0.024			
3	1			0.007		
4	1				0.013	
5	1					0.053
6	2	0.030	0.024			
7	2	0.035		0.022		
8	2	0.022			0.033	
9	2	0.084				0.076
10	2		0.042	0.017		
11	2		0.023		0.014	
12	2		0.029			0.028
13	2			0.045	0.037	
14	2			0.016		0.022
15	2				0.015	0.017
16	3	0.033	0.022	0.021		
17	3	0.015	0.053		0.039	
18	3	0.035	0.023			0.027
19	3	0.018		0.029	0.045	
20	3	0.018		0.029		0.017
21	3	0.017			0.019	0.018
22	3		0.025	0.018	0.033	
23	3		0.031	0.020		0.028
24	3		0.022		0.015	0.018
25	3			0.026	0.031	0.027
26	4	0.025	0.019	0.023	0.021	
27	4	0.023	0.018	0.022		0.036
28	4	0.020	0.021		0.019	0.024
29	4	0.027		0.015	0.012	0.023
30	4		0.020	0.026	0.014	0.020
31	5	0.024	0.019	0.022	0.016	0.021

Table 2. The mean absolute error (MAE) on examination set for determination of the concentrations of salts in different combinatorial classes by neural network Method 2. The marked results are best among all the methods

Class number	Number of salts	NaCI	NH_4Br	Li_2SO_4	KNO_3	CsI
1	1	0.0038				
2	1		0.0004			
3	1			0.0002		
4	1				0.002	
5	1					0.001
6	2	0.039	0.021			
7	2	0.022		0.023		
8	2	0.030			0.046	
9	2	0.108				0.078
10	2		0.055	0.022		
11	2		0.029		0.019	
12	2		0.023			0.027
13	2			0.037	0.042	
14	2			0.024		0.017
15	2				0.025	0.019
16	3	0.037	0.024	0.026		
17	3	0.025	0.118		0.080	
18	3	0.036	0.015			0.040
19	3	0.037		0.023	0.032	
20	3	0.031		0.021		0.023
21	3	0.032			0.027	0.015
22	3		0.016	0.020	0.025	
23	3		0.028	0.016		0.029
24	3		0.027		0.060	0.017
25	3			0.018	0.029	0.024
26	4	0.031	0.022	0.021	0.017	
27	4	0.026	0.018	0.024		0.038
28	4	0.025	0.021		0.020	0.039
29	4	0.030		0.019	0.020	0.025
30	4		0.018	0.021	0.016	0.017
31	5	0.026	0.019	0.019	0.016	0.021

Table 3. The mean absolute error (MAE) on examination set for determination of the concentrations of salts in different combinatorial classes by PLS method on initial data. The marked results are best among all the methods

Class number	Number of salts	NaCl	NH_4Br	Li_2SO_4	KNO_3	CsI
1	1	0.0028				
2	1		0.000015			
3	1			0.0000092		
4	1				0.00002	
5	1					0.000016
6	2	0.024	0.024			
7	2	0.030		0.026		
8	2	0.016			0.042	
9	2	0.089				0.083
10	2		0.038	0.018		
11	2		0.020		0.016	
12	2		0.019			0.029
13	2			0.054	0.041	
14	2			0.022		0.010
15	2				0.030	0.013
16	3	0.045	0.017	0.043		
17	3	0.017	0.045		0.030	
18	3	0.029	0.013			0.023
19	3	0.017		0.022	0.022	
20	3	0.025		0.028		0.017
21	3	0.021			0.017	0.023
22	3		0.014	0.014	0.022	
23	3		0.024	0.022		0.026
24	3		0.024		0.025	0.014
25	3			0.017	0.033	0.021
26	4	0.031	0.017	0.027	0.023	
27	4	0.031	0.016	0.022		0.032
28	4	0.018	0.015		0.019	0.026
29	4	0.037		0.020	0.013	0.025
30	4		0.018	0.021	0.019	0.017
31	5	0.040	0.027	0.037	0.025	0.028

Table 4. The mean absolute error (MAE) on examination set for determination of the concentrations of salts in different combinatorial classes by PLS method on preprocessed data. The marked results are best among all the methods

Class number	Number of salts	NaCl	NH_4Br	Li_2SO_4	KNO_3	CsI
1	1	0.0026				
2	1		0.000021			
3	1			0.000013		
4	1				0.000018	
5	1					0.000042
6	2	0.024	0.024			
7	2	0.030		0.026		
8	2	0.016			0.041	
9	2	0.090				0.065
10	2		0.038	0.018		
11	2		0.020		0.014	
12	2		0.020			0.028
13	2			0.054	0.042	
14	2			0.021		0.011
15	2				0.028	0.012
16	3	0.046	0.017	0.042		
17	3	0.017	0.043		0.046	
18	3	0.029	0.013			0.023
19	3	0.017		0.022	0.023	
20	3	0.024		0.028		0.017
21	3	0.020			0.016	0.023
22	3		0.015	0.015	0.022	
23	3		0.024	0.022		0.026
24	3		0.024		0.026	0.013
25	3			0.018	0.034	0.021
26	4	0.031	0.017	0.027	0.022	
27	4	0.030	0.016	0.021		0.032
28	4	0.018	0.015		0.019	0.026
29	4	0.037		0.020	0.013	0.025
30	4		0.017	0.020	0.018	0.016
31	5	0.028	0.019	0.025	0.018	0.021

CONSLUSIONS

Comparative analysis of different methods for solving the complex inverse problem of determining the types and individual concentrations of salts in 5-component aqueous solutions by Raman spectra, within the "experiment-based" approach by the two bands of the Raman spectrum of aqueous solutions—the low-frequency band (280–1830 cm^{-1}) and the valence band of water (2700–3900 cm^{-1})—has been performed.

Use of NN to solve this problem turned out to be more efficient within the framework of Method 1 (identification and determination of individual concentrations of all salts in a single stage). Worse results given by Method 2 (identification of the salts by a single common NN with subsequent determination of their concentrations by another NN trained specially for the given component composition) may be explained by the decrease in the ratio of the number of samples and the number of input features when working with a specific component composition of the solution. Data aggregation reduces this effect, but it does not eliminate it.

It has been demonstrated that the best results in the cases when the non-linearity of the problem is not too strong can be achieve by the PLS-regression method. In cases when the non-linearity is more significant, transformation of transmission spectra to absorption spectra allows improving the quality of the IP solution by the PLS-regression method. In the most non-linear cases, the neural network method of the problem solution remains more efficient.

REFERENCES

1. Terekhov, S.A., "Direct, inverse and combined problems in complex engineered system modeling by artificial neural networks", *Proc. SPIE AeroSense Conference, Orlando, Florida, 1997*, Proc. SPIE, vol. 3077, paper 71.
2. Clark, John W., "Neural networks: new tools for modeling and data analysis in science", in *Scientific Applications of Neural Nets, Lecture Notes in Physics*, Volume 522, 1999, pp. 1–96.
3. Esbensen, K.H., Guyot, D., Westad, F., and Houmoller, L.P. *Multivariate Data Analysis—In Practice: An Introduction to Multivariate Data Analysis and Experimental Design*, 5th ed., CAMO Software AS, 2006.
4. Rumondor, A.C. and Taylor, L.S., Application of partial least-squares (PLS) modeling in quantifying drug crystallinity in amorphous solid dispersions, *Int. J. Pharm.*, 2010, vol. 398, nos. 1–2, pp. 155–160.
5. Nguyen, Danh V. and Rocke, D.M., Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics*, 2002, vol. 18, no. 1, pp. 39–50.
6. Tan, F., Feng, X., Fang, Z., Li, M., Guo, Y., and Jiang, L., Prediction of total antioxidant capacity of fruit juices using FTIR spectroscopy and PLS regression, *Amino Acids*, 2007, vol. 33, no. 4, pp. 669–675.
7. Li, Xiaohong, Gill, Ryan, Cooper, Nigel, G.F., Yoo, Jae Keun, and Datta, Susmita, Modeling microRNA-mRNA interactions using PLS regression in human colon cancer, *BMC Med. Genomics,* 2011, vol. 4, p. 44.
8. Baldwin, S.F. and Brown, C.W., "Detection of ionic water pollutants by laser excited Raman spectroscopy", *Water Res.*, 1972, vol. 6, pp. 1601–1604.
9. Rudolph, W.W. and Irmer, G., "Raman and infrared spectroscopic investigation on aqueous alkali metal phosphate solutions and density functional theory calculations of phosphate-water clusters", *Appl. Spectroscopy*, 2007, vol. 61, no. 12, pp. 274A–292A.
10. Furic, K., Ciglenecki, I., and Cosovic, B., "Raman spectroscopic study of sodium chloride water solutions", *J. Mol. Structure*, 2000, vol. 6, pp. 225–234.
11. Dolenko, T.A., Churina, I.V., Fadeev, V.V., and Glushkov, S.M., "Valence band of liquid water Raman scattering: some peculiarities and applications in the diagnostics of water media", *J. Raman Spectroscopy*, 2000, vol. 31, pp. 863–870.
12. Burikov, S.A., Dolenko, T.A., Fadeev, V.V., and Sugonyaev, A.V., "New opportunities in the determination of inorganic compounds in water by the method of Laser Raman Spectroscopy", *Laser Phys.*, 2005, vol. 15, no. 8, pp. 1–5.
13. Burikov, S.A., Dolenko, T.A., and Fadeev, V.V., "Identification of inorganic salts and determination of their concentrations in water solutions from the Raman valence band using artificial neural networks", *Pat. Rec. Image Analysis*, 2007, vol. 17, no. 4, pp. 554–559.
14. Burikov, S.A., Dolenko, T.A., Dolenko, S.A., and Persiantsev, I.G., Neural network solution of inverse problem of identification and determination of partial concentrations of inorganic salts in the multicomponent aqueous solution, in *Neuroinformatics-2010 XII*, Moscow: MEPhI, 2010, vol. 2, pp. 100–110 (in Russian).
15. Burikov, S.A., Dolenko, T.A., Dolenko, S.A., and Persiantsev, I.G., Application of artificial neural networks to solve problems of identification and determination of concentration of salts in multi-component water solutions by Raman spectra, *Opt. Mem. Neural Networks (Inform. Optics)*, vol. 19, no. 2, pp. 140–148.

16. Burikov, S.A., Dolenko, S.A., Dolenko, T.A., and Persiantsev, I.G., Application of artificial neural networks to solve problems of identification and determination of concentration of salts in multi-component water solutions by Raman spectra, *Opt. Mem. Neural Networks (Inform. Optics)*, 2010, vol. 19, no. 2, pp. 140–148.
17. Dolenko, S.A., Burikov, S.A., Dolenko, T.A., and Persiantsev, I.G., Adaptive methods for solving inverse problems in laser Raman spectroscopy of multi-component solutions, *Pat. Rec. Image Analysis*, 2012, vol. 22, no. 4, pp. 551–558.
18. Dolenko, S., Burikov, S., Dolenko, T., Efitorov, A., and Persiantsev, I., Methods of input data compression in neural network solution of inverse problems of spectroscopy of multi-component solutions, *11th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2003)*, Samara: IPSI RAS, vol. 2, pp. 541–544.
19. Gerdova, I.V., Churina, I.V., Dolenko, S.A., Dolenko, T.A., Fadeev, V.V., and Persiantsev, I.G., "New opportunities in solution of inverse problems in laser spectroscopy due to application of artificial neural networks", *Proceedings of SPIE—The International Society for Optical Engineering*, *SPIE, the International Society for Optical Engineering (Bellingham, WA, United States)*, 2012, vol. 4749, pp. 157–166.