CLUSTER-BASED APPROACH TO THE SOLUTION OF THE INVERSE PROBLEM OF DIAGNOSTICS OF MUTI-COMPONENT SOLUTIONS BY ADAPTIVE METHODS

Alexander Efitorov[1], Sergey Burikov[1,2], Tatiana Dolenko[1,2], Kirill Laptinskiy[1,2], and Sergey Dolenko[1]

1. D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, Russia; dolenko@srd.sinp.msu.ru
2. Physical Department, M.V.Lomonosov Moscow State University, Moscow, Russia; tdolenko@mail.ru

Novel methods of water chemistry control are demanded in ecology and industry, because the classical approaches usually require direct contact with studied samples and a long time for analysis. The authors (1) suggested the method of identification and determination of concentration of inorganic salts dissolved in water by Raman spectroscopy and artificial neural networks (ANN), particularly by multilayer perceptron (MLP) (2), at the example of solution of this inverse problem (IP) for five salts. Due to the complexity of the object, there is no adequate physical model that would allow obtaining the dependence of the water Raman spectrum on concentrations of the dissolved salts numerically. Therefore, authors had to apply machine learning methods to solve the IP within the “experiment-based” approach (3). Further it was demonstrated (4), that dividing the data array into groups according with the chemical composition and determining concentration within each group by a partial least squares (PLS) algorithm (5) provided better accuracy than ANN, trained either on the full or on a divided data array. Also, in (6), two approaches to data array division were compared: supervised division according to the types of the dissolved salts, and unsupervised division performed by a number of machine learning clustering algorithms. It was demonstrated that the second approach leads to deterioration of the results of IP solution, but not a very dramatic one, particularly if the division is implemented by a Kohonen self-organization map (SOM) (7).

In this paper[*], the studied inverse problem is determination of ionic composition of a solution of inorganic salts: $MgSO_4$, $Mg(NO_3)_2$, LiCl, $LiNO_3$, NH_4F, $(NH_4)2SO_4$, $KHCO_3$, KF, $NaHCO_3$, and NaCl. The total concentration of salts in the solutions changed in the range from 0 to 1.5 M, concentration of each salt changed in the range from 0 to 1.5 M with concentration step 0.15-0.25 M. A total of 4445 Raman spectra of water solutions were recorded; each spectrum had 1824 channels in the Raman shift range 565…4000 cm^{-1}. The procedure of data preparation and pre-processing was described in (1). Next, to reduce the input dimension of the problem, 213 significant features were selected by the method of ANN weight analysis (8). Obviously, it was impossible to form adaptive models on all subsets based of types of the dissolved salts, because this approach would yield 210 arrays containing each 4 spectra on the average. Thereby, SOM was used for division of the total data set by means of data clustering. The regression problem was solved by PLS and MLP ANN within each cluster.

Figure 1 presents the dependence of the mean absolute error of the IP solution on the number of clusters; all the presented results are obtained on the samples of an external validation set. Apparently, there are strong (compared with solutions of five salts) nonlinear interactions in the water solutions, due to which the ANN with a large number of neurons demonstrates better results than PLS. Unfortunately, MLP requires a large training dataset; therefore, the best result was demonstrated on the full undivided data array. A similar behaviour was observed earlier for the IP of five salts (4).

[*] This study has been performed at the expense of a grant of the Russian Science Foundation, project no.14-11-00579.

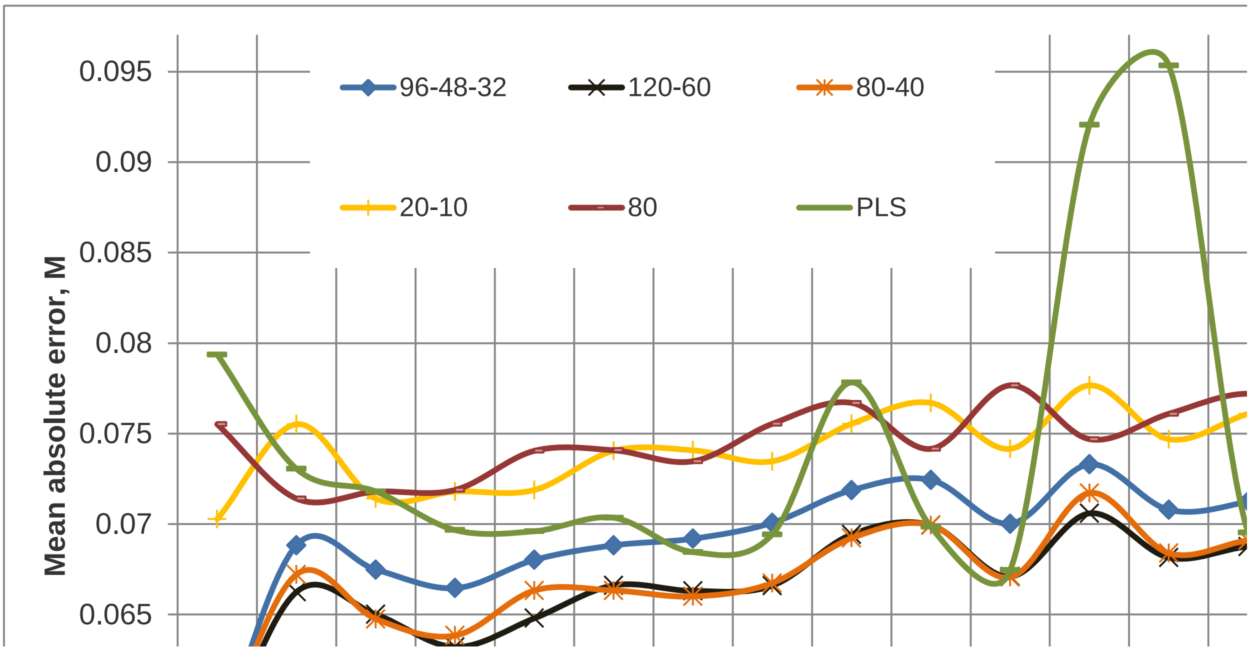

Figure 1: The dependence of mean absolute error (averaged by ions) of IP solution by PLS (green line) and MLP (all the others; number of neurons in hidden layers is given in the legend) regression models on number of clusters.

REFERENCES

1 Burikov S A, S A Dolenko, T A Dolenko & I G Persiantsev, 2010. Application of Artificial Neural Networks to solve problems of identification and determination of concentration of salts in multi-component water solutions by Raman spectra. Optical Memory and Neural Networks (Information Optics), 19(2): 140-148

2 Haykin S, 1999. Neural Networks - A Comprehensive Foundation, 2nd Edition (Pearson Education) 823 pp.

3 Gerdova I V, S A Dolenko, T A Dolenko, I G Persiantsev, V V Fadeev & I V Churina, 2002. New opportunity solutions to inverse problems in laser spectroscopy involving artificial neural networks. Izv.AN SSSR Seriya Fizicheskaya, 66(8): 1116-1124

4 Efitorov A O, S A Burikov, T A Dolenko, I G Persiantsev & S A Dolenko, 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. Optical Memory and Neural Networks (Information Optics), 24(2): 93-101

5 Wold S, P Geladi, K Esbensen & J Ohman, 1987. Multi-way principal components- and PLS-analysis. Journal of Chemometrics, 1(1): 41-56

6 Dolenko S, A Efitorov, S Burikov, T Dolenko, K Laptinskiy & I Pesiantsev, 2015. Neural Network approaches to solution of the inverse problem of identification and determination of the ionic composition of multi-component water solutions. Communications in Computer and Information Science, 517: 109-118.

7 Kohonen T, 2001. Self-Organizing Maps, 3rd ed. (Springer International Publishing) 501 pp.

8 Efitorov A, S Burikov, T Dolenko, K Laptinskiy & S Dolenko, 2015. Significant feature selection in Neural Network solution of an inverse problem in spectroscopy. Procedia Computer Science, 66: 93-102