Adaptive Methods for Solving Inverse Problems in Laser Raman Spectroscopy of Multi-Component Solutions[1]

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

[a]*Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, Leninskie Gory, Moscow, 119991 Russia*
[b]*Department of Physics, Lomonosov Moscow State University, Leninskie Gory, Moscow, 119991 Russia*
e-mail: dolenko@srd.sinp.msu.ru, burikov@lid.phys.msu.ru, tdolenko@lid.phys.msu.ru, ipers@srd.sinp.msu.ru

Abstract—This study provides comparative analysis of approaches connected with application of neural network based algorithms for efficient solution of pattern recognition problem (inverse problem with discrete output) combined with solution of inverse problem with continuous output. The analysis is performed at the example of the problem of identification and determination of concentrations of inorganic salts in multi-component aqueous solutions by Raman spectrum.

INTRODUCTION

It is well known that artificial neural networks (ANN) are a class of mathematical algorithms that have demonstrated high efficiency in the solution of approximation, prediction, evaluation, classification and pattern recognition problems. ANN are also widely used for solution of inverse problems (IP), where such their properties as learning by examples, high noise immunity, and immunity to contradictory data, play a special role [1, 2]. In this study, development of methods of solving such problems is performed at the example of the complex IP of identification and determination of partial concentrations of inorganic salts in multi-component aqueous solution by Raman spectra.

The problem of determination of concentrations of substances dissolved in water is very important for oceanology, ecological monitoring, and control of mineral, technical and waste waters. It is required to solve this IP in non-contact express mode with reasonable accuracy.

Method of Raman spectroscopy complies with these requirements. Principle opportunity of use of water Raman spectra for diagnostics of solutions arises from high sensitivity of their characteristics to type and concentration of salts dissolved in water. In [3, 4] it is suggested to use Raman spectra of complex ions (such as bands of anions NO_3^-, SO_4^{2-}, PO_4^{3-}, CO_3^{2-} near 1000 cm^{-1}) for determination of type and concentrations of salts in water. The type of anion can be

determined by the position of the corresponding band, its concentration—by the band intensity. But this method is used only for analysis of substances having their own Raman bands, i.e., for salts with complex anions. The authors of [5–8] elaborated a method of determination of concentration of dissolved salts by water Raman valence band. In [5, 6], concentration of one salt in the solution was determined using water Raman valence band. In [7, 8], a method of determination of partial concentrations of several salts in multi-component solutions by water Raman valence band, was suggested and elaborated. It was possible because of application of ANN for solution of this IP.

1. METHODOLOGY

Training of ANN requires a representative dataset, i.e., a set of data that would reflect all characteristic aspects of behavior of the object. This dataset can be obtained in different ways. So, there are the following methodical approaches to solution of IP using ANN [9].

(a) "Model-based" approach. If an adequate analytical model of solution of the direct problem is available, it can be used for generation of arrays of data with necessary representativity. The main disadvantage of this approach is that elaboration and realization of an adequate model is often impossible or very difficult. This approach is unacceptable in our case. Because of complexity of the object, there is no adequate model based on physical reasons that would allow obtaining the dependence of water Raman spectrum on concentration of dissolved salts, especially taking into account their non-linear interaction.

(b) "Experiment-based" approach. The data used for ANN training are obtained in experiment. This

---

[1] The article was translated by the authors.

approach does not require a model to be available, and it allows taking non-linear properties of the object into account. However, obtaining a representative data set can be a non-trivial experimental problem. In this study, 8695 experimental spectra for 4268 different solutions were obtained for realization of the "experiment-based" approach.

(c) "Quasi-model" approach. If no adequate physically grounded analytical model of solution of the direct problem is available, one can replace it by a parametrical "quasi-model" based on experimental data. This model formally describes the dependence of the observed data on the sought-for parameters. It was shown in [7] that because of non-linear interactions in the solutions, the approach involving the simplest linear quasi-model is unusable for determination of concentrations of salts in multi-component solutions.

In order to construct more complicated quasi-models, which can describe the desired dependence better, one can use more efficient adaptive methods for construction of models—for example, Group Method of Data Handling (GMDH) [10] or some types of ANN [11, 12].

2. EXPERIMENT

Excitation of Raman spectra was performed by argon laser (wavelength 488 nm, output power 450 mW). Raman spectra of water solutions of inorganic salts were measured in $90°$ geometry using monochromator (resolution 2 cm^{-1}) and CCD-camera (Jobin Yvon). Spectra were measured in two regions: 300—2300 and 2300—4000 cm^{-1} for every sample. The temperature of samples during experiment was stabilized at $22.0 \pm 0.2°C$.

The objects of research were water solutions of the salts NaCl, NH_4Br, Li_2SO_4, KNO_3, CsI. Concentration of every salt in the solution was changed from 0 to 2.5 M (step 0.2—0.25 M). In Fig. 1, there are Raman spectra of water and water solutions in the region 300—1800 cm^{-1}. One can observe bands of valence vibrations of anions SO_4^{2-} and of anions NO_3^{-} (Fig. 1a). Their intensities depend on concentrations of the corresponding salts (Fig. 1b).

In Fig. 2, it can be seen that water Raman valence band depends on salt types (Fig. 2a) and on their concentrations (Fig. 2b): under increase of concentration of salts, water Raman valence band (2600—4000 cm^{-1}) shifts towards high frequencies, its half-width decreases, the intensity of its high-frequency part increases, and the intensity of the low-frequency part decreases. The changes in position and shape of water valence band also depend on type and concentration of salt [6].

3. RESULTS

To train ANN, the whole data array was divided into training, test and examination sets (with 70 : 20 : 10 ratio). The training set was used to train networks. The test set was used for regular testing during training, to interrupt training at the minimum error on test set and thus to prevent overtraining. All the results presented below were obtained on the out-of-sample (examination) data set, thus providing impartial assessment of the results. Also, the results on all the three data sets were similar, what gave evidence of sufficient representativity of the examination data set.

3.1. "Experiment-based" Approach

Within the "experiment-based" approach, two types of network architecture were used.

For the first type, an ANN had five outputs, according to the maximal number of components in a solution. Each output corresponded to one of the considered salts, and its desired value was equal to concentration of this salt in the solution. Two NN models were used to solve the considered problem: perceptrons [11] with three hidden layers, and General regression neural networks (GRNN) [12]. Better results were shown by perceptron with three hidden layers.

For the second type of network architecture, five separate similar networks were used, each having one output. The desired output of each network was equal to concentration of the corresponding salt in the solution. Here also two different types of adaptive algorithms were tested: perceptrons with three hidden layers, and GMDH [10]. In this case, perceptrons also demonstrated better results.

Table 1 compares the results of solution of the considered problem by perceptrons.

Computational experiments (Exps. 1 and 2) [7, 8] were performed for diagnostics of three-component solutions using only valence band of the spectrum. Significantly better results (smaller error values) obtained in Exps. 3, 4 [13, 14] in a wider range of concentrations and for five-component solutions, can be explained not only by higher spectral resolution and better quality of experimental spectra, but also by the fact that the low-frequency region of spectrum (Fig. 1), where bands of complex ions are present, was taken into consideration. Exps. 3—7 were based on the same experimental material.

In Exp. 5, five single-output networks were used instead of one five-output network. As can be seen from the Table 1, there is an improvement in the results of Exp. 5 compared to Exp. 4, but it is not very significant.

Exps. 3—5 used selection of input features by absolute value of standard deviation (st.dev, StD) of spectrum intensity in each channel, as this StD is proportional to the entropy and thus to the volume of information brought by this channel; 704 features selected corresponded to channels with StD exceeding some threshold. In Exp. 6, GMDH [10] was used to determine subsets of this set of features most significant for determination of concentration of each salt. Union of
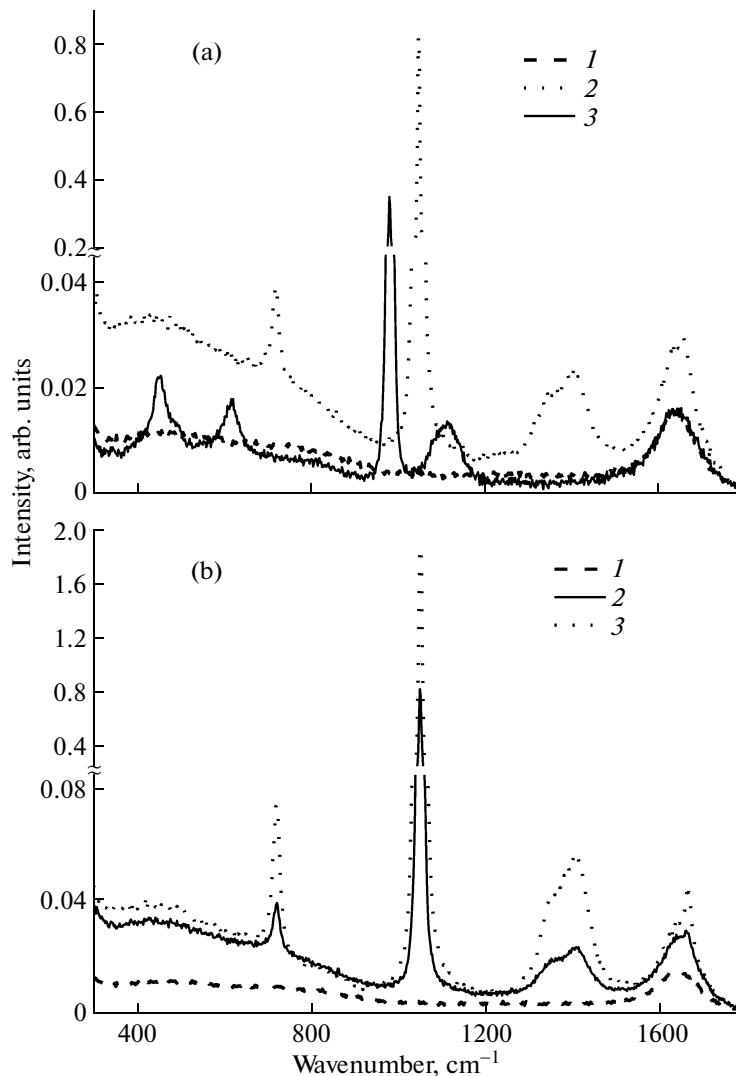
Fig. 1. Low-frequency region of spectra of water solutions of different salts with complex anions: (a) *1*—distilled water; *2*—KNO$_3$, 1 M; *3*—Li$_2$SO$_4$, 1 M, (b) *1*—distilled water; *2*—KNO$_3$, 1 M; *3*—KNO$_3$, 3 M.

these subsets gave 314 features used in Exp. 6. Comparing the results of Exps. 4 and 6 (Table 1), one can see that on the average, the reduction of the number of input features did not result in error reduction.

The main shortcoming of the algorithm of feature selection described above is that some part of the information brought by any channel (input feature) can be useless for solution of the problem. In Exp. 7, feature selection was performed based on the values of cross-correlation (CC) and cross-entropy (CE). The average value and the StD were computed over the whole array of CC of an output feature with each input feature. Input features whose CC exceeded the sum of the average CC and its StD, were selected. This procedure has been repeated for each output and for CE instead of CC. Union of the thus obtained 10 sets of input features gave 1134 inputs; it turned out that some of the 704 features selected by the first method were

not included in this number. Comparison of the results of Exp. 7 to those of Exp. 4 in Table 1 shows that while in general the errors became smaller, the decrease is not very significant.

3.2. "Quasi-Model" Approach: Construction of Quasi-Models

Not very high accuracy of solution of an inverse problem using the "experiment-based approach" is often reasoned by the unfavorable ratio of the number of input variables of the problem (in this study, 704 in Exps. 3–5) and the number of patterns in the training set (in this study, 8229 in training and test sets together). One can overcome it in two ways: by further reduction of the number of input variables or by increase of number of patterns.
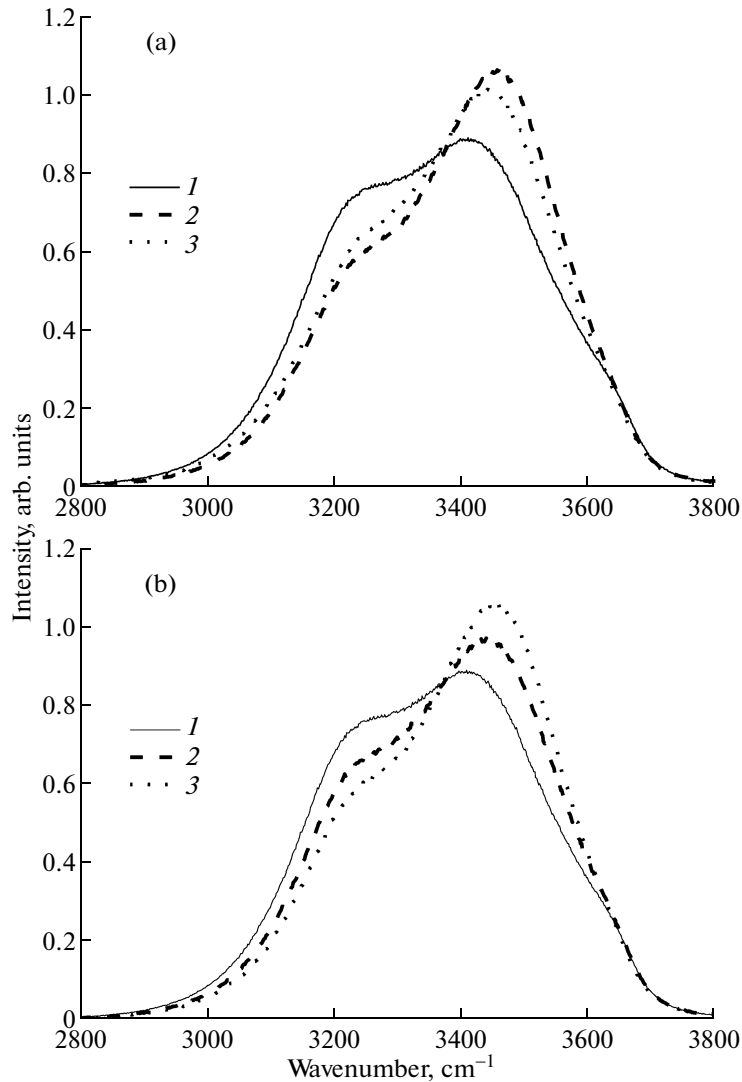
Fig. 2. Water Raman valence bands: (a) *1*—distilled water; *2*—CsI, 2 M; *3*—NaCl, NH$_4$Br, Li$_2$SO$_4$—0.4 M, KNO$_3$ and CsI—0.6 M, (b) *1*—distilled water; *2*—CsI, 1 M; *3*—CsI, 2 M.

When realization of additional experiments is impossible, too expensive, or takes unacceptably much time, one can try to get additional patterns by interpolation using a quasi-model—a parametrical or an adaptive model of solution of the direct problem. However, one should realize that if the quasi-model is not adequate enough (i.e. if it does not describe the solution of the direct problem sufficiently well), use of such "quasi-model" approach can lead to deterioration of quality of IP solution instead of its improvement. In particular, this was the case for a simple linear combination quasi-model applied to the data used in this study [7].

In this paper, three types of quasi-models were under consideration: those based on perceptron, GRNN, and GMDH.

To estimate applicability of the "quasi-model" approach for this problem, first it was necessary to choose the best of the constructed quasi-models (the one providing minimal error of solution of the direct problem on the examination data set).

In Table 2, one can see statistical indexes, their minimal, maximal, mean values and standard deviations over all the 704 simulated channels. Let us accept the following notion:

y—the estimate of the value of the output variable made by the model (neural network);

d—the desired value of this output variable;

$\bar{d}$—the mean value of the desired output over the whole considered data set;

N—the number of patterns in the considered data set.

Then the statistical indexes used in Tables 2–4 are the following.

Table 1. Mean absolute error for determination of concentrations of salts in multi-component solutions by Raman spectra, within the "experiment-based" approach (M, on examination data set)

Exp. no.	1	2	3	4	5	6	7
Reference	[7]	[8]	[13, 14]		This study		
Components	3	3	5				
Range	0.7 M	1 M	2.5 M				
Bands		Valence only		Both bands			
Inputs	200	200	704	704	704	314	1134
Selected by	none	none	st. dev	st. dev	st. dev	GMDH	CC, CE
Nets	1	1	1	1	5	1	1
NaCl	0.07	0.07	0.047	0.029	0.029	0.031	0.028
NH_4Br	0.06	0.11	0.029	0.024	0.024	0.025	0.022
Li_2SO_4	–	–	0.040	0.020	0.019	0.018	0.019
KNO_3	–	–	0.046	0.019	0.017	0.018	0.018
CsI	–	–	0.032	0.023	0.023	0.023	0.024
KI	0.05	0.12	–	–	–	–	–

The *coefficient of multiple determination R squared* (R^2) is computed by the formula:

$$R^2 = 1 - \frac{\sum (d - y)^2}{\sum (d - \bar{d})^2}.$$

This index compares the precision of the tested model to that of the trivial benchmark model whose estimate is the average value of the desired output over the whole considered data set. For the absolutely correct estimate, the value of R^2 squared is equal to 1; for an estimate worse than the trivial one, R^2 squared becomes negative. For many types of problems, R^2 squared is the most substantive universal criterion for model quality estimation. The value of R^2 is dimensionless.

The *mean relative error* (MRE) is computed by the formula:

$$COO = \frac{1}{N} \sum \frac{|d - y|}{d} \times 100\%.$$

This index is also dimensionless, and it is usually expressed in per cent.

Table 2. Statistical indexes of construction of quasi-models based on perceptron and GRNN (on examination data set)

Index, model	min	max	mean	st. dev.
R^2, perceptron	0.681	0.983	0.933	0.048
R^2, GRNN	0.768	0.974	0.933	0.041
MRE, %, perceptron	0.5	272.9	9.9	27.0
MRE, %, GRNN	0.4	75.0	7.1	11.6

Summation in both formulas is performed over all the patterns of the considered data set (from 1 to N).

Remember that a quasi-model in our case is a model dependence of intensity in every spectral channel on concentrations of components in the solution. Indexes in Table 2 are shown for quasi-models based on perceptron and GRNN. One can see that mean indexes in Table 2 are close, and there is no reason to prefer either of examined quasi-models at this stage.

Construction of a quasi-model based on GMDH is carried out separately for every channel, and therefore this routine is much more labor-consuming. That is why such construction was first performed for four channels where the quasi-models based on GMDH and perceptron demonstrated results of different quality (bad or good). Comparison of the results for all the three types of models is presented in Table 3.

One can see that there is no pronounced advantage of one or another quasi-model. For this reason, further studies of application of the "quasi-model" approach were performed for quasi-models based on perceptron and GRNN only. The quasi-model based on GMDH is interesting according to two aspects: as a method of selection of significant input variables for solution of IP (Table 1, Exp. 6), and as a method to obtain relatively simple equations for solution of direct problem, which can be an object of independent analysis. However, studies in the second direction were out of scope of this study.

3.3. "Quasi-Model" Approach: Implementation

The data for the "quasi-model" approach were prepared as follows.

Table 3. Comparison of the results of construction of quasi-models based on perceptron, GRNN and GMDH for some spectral channels (on examination data set)

Index, model	Channel 1 951.4 cm^{-1}	Channel 48 1047.3 cm^{-1}	Channel 185 1672.2 cm^{-1}	Channel 590 3494.1 cm^{-1}
R^2, perceptron	0.830	0.951	0.726	0.983
R^2, GRNN	0.786	0.974	0.768	0.970
R^2, GMDH	0.665	0.963	0.677	0.978
MRE, %, perceptron	24.4	198.5	8.8	0.75
MRE, %, GRNN	23.0	73.4	7.5	0.97
MRE, %, GMDH	18.3	59.4	8.9	0.79

Table 4. Some statistical indexes for determination of salts concentration in multi-component solutions by Raman spectra using neural networks within "experiment-based" and "quasi-model" approaches. Specification of data sets: EE—experimental examination, QE—quasi-model examination. Specification of approaches: Exp—"experiment-based," QMP and QMGR—"quasi-model" based on perceptron and GRNN, respectively

Approach	Set	NaCl	NH$_4$Br	Li$_2$SO$_4$	KNO$_3$	CsI
				R^2		
Exp	EE	0.990	0.993	0.996	0.996	0.994
QMP	EE	0.969	0.985	0.983	0.986	0.987
QMGR	EE	0.980	0.986	0.992	0.991	0.989
QMP	QE	0.989	0.995	0.990	0.991	0.996
QMGR	QE	0.990	0.992	0.991	0.991	0.992
				COO, %		
Exp	EE	9.2	5.8	5.6	4.3	5.9
QMP	EE	17.4	9.3	10.4	8.2	9.3
QMGR	EE	12.5	8.8	7.7	6.8	7.5
QMP	QE	10.4	7.2	8.6	8.9	6.2
QMGR	QE	11.3	9.5	9.5	9.5	9.1

To create "quasi-model" data sets, a grid with concentration step 0.15 M for all salts was chosen. Only spectra with total concentration less than 2.5 M were considered. There were 535130 such spectra with different salt concentrations (i.e., more than 10 times as many as experimental spectra with different concentrations). The obtained array of 53130 "quasi-model" spectra was randomly divided into training, test and examination data sets in the ratio 70 : 20 : 10. After that, training and test sets were complemented by training and test sets from experimental array. Examination sets were left separate. Thus for application of "quasi-model" approach the following data sets were used: training (43593 patterns), test (12454 patterns), examination quasi-model (5313 patterns) and examination experimental (914 patterns).

Two such complete arrays were obtained: one by generation of "quasi-model" spectra using the quasi-model based on perceptron and another one based on GRNN.

The obtained arrays were used to train identical perceptrons with the same architecture and training parameters as those used for solution of IP within the "experiment-based" approach (Table 1, Exp. 4).

Table 4 compares the results obtained within the "experiment-based" approach on the examination (experimental examination) data set and the results obtained within the "quasi-model" approach on quasi-model examination and experimental examination data sets, for two types of "quasi-model"—based on perceptron and based on GRNN.

The obtained results allow one to make the following conclusions:

1. The "quasi-model" approach deceived expectations in comparison with the "experiment-based" approach. In all cases the results on experimental examination set obtained within the "quasi-model" approach were worse or considerably worse than the initial results obtained within the "experiment-based" approach. This indicates low adequacy of the used quasi-models.

2. The quasi-model based on GRNN in all cases produced higher results on the experimental examination set than the quasi-model based on perceptron.

3. The results obtained within the "quasi-model" approach on quasi-model examination set are comparable with those obtained within the "experiment-based" approach on experimental examination set. This is a revelation of higher representativity of data sets in the "quasi-model" approach.

CONCLUSIONS

The complex inverse problem of identification of salts and determination of their partial concentrations in 5-component water solution by Raman spectra was solved within the "experiment-based" approach using two bands of Raman spectrum—low-frequency band ($300–1830$ cm^{-1}) and valence band ($2700–3900$ cm^{-1}), as well as using only water Raman valence band. When both bands of experimental Raman spectra were used as input data, the obtained values of the error of determination of concentration on the examination dataset were low enough: the mean absolute error was from 0.02 to 0.03 M in the concentration range from 0 to 2.5 M.

Attempts of further reduction of the error included: using a separate single-output network for each salt instead of a common 5-output network; using the "quasi-model" approach to increase the number of patterns; use of GMDH for further reduction of the number of input features; use of a combination of cross-correlation and cross-entropy as an alternative method of feature selection. However, error reduction brought by all the applied methods was insignificant.

Further studies should include use of more sophisticated feature selection or feature extraction methods, and elaboration of more adequate quasi-models.

ACKNOWLDGMENTS

REFERENCES

1. S. A. Terekhov, "Direct, Inverse, and Combined Problems in Complex Engineered System Modeling by Artificial Neural Networks," in *Proc. SPIE AeroSense Conf.* (Orlando, FL, Apr. 21–24, 1997), Vol. 3077.

2. S. A. Dolenko, T. A. Dolenko, I. G. Persiantsev, V. V. Fadeev, and S. A. Burikov, "The Way to Solve the Spectroscopy Inverse Problems by Means of Neural Networks," Neirokomp'yut.: Razrab. Prim., Nos.1–2, 89–97 (2005).

3. S. F. Baldwin and C. W. Brown, "Detection of Ionic Water Pollutants by Laser Excited Raman Spectroscopy," Water Res. **6**, 1601–1604 (1972).

4. W. W. Rudolph and G. Irmer, "Raman and Infrared Spectroscopic Investigation on Aqueous Alkali Metal Phosphate Solutions and Density Functional Theory Calculations of Phosphate-Water Clusters," Appl. Spectroscopy **61** (12), 274A–292A (2007).

5. K. Furic, I. Ciglenecki, and B. Cosovic, "Raman Spectroscopic Study of Sodium Chloride Water Solutions," J. Molec. Struct. **6**, 225–234 (2000).

6. T. A. Dolenko, I. V. Churina, V. V. Fadeev, and S. M. Glushkov, "Valence Band of Liquid Water Raman Scattering: Some Peculiarities and Applications in the Diagnostics of Water Media," J. Raman Spectroscopy **31**, 863–870 (2000).

7. S. A. Burikov, T. A. Dolenko, V. V. Fadeev, and A. V. Sugonyaev, "New Opportunities in the Determination of Inorganic Compounds in Water by the Method of Laser Raman Spectroscopy," Laser Phys. **15** (8), 1–5 (2005).

8. S. A. Burikov, T. A. Dolenko, and V. V. Fadeev, "Identification of Inorganic Salts and Determination of Their Concentrations in Water Solutions from the Raman Valence Band Using Artificial Neural Networks," Pattern Recogn. Image Anal. **17** (4), 554–559 (2007).

9. I. V. Gerdova, I. V. Churina, S. A. Dolenko, T. A. Dolenko, and V. V. Fadeev, "New Opportunities in Solution of Laser Spectroscopy Inverse Problems Using Artificial Neural Networks," Izv. Akad. Nauk. Ser. Phys. **66** (8), 1116–11234 (2002)

10. H. R. Madala and A. G. Ivakhnenko, *Inductive Learning Algorithms for Complex Systems Modeling* (CRC Press, 1994).

11. M. H. Hassoun, *Fundamentals of Artificial Neural Networks* (MIT Press, Cambridge, 1995).

12. D. Specht, "A General Regression Neural Network," IEEE Trans. Neural Networks **2** (6), 568–576 (1991).

13. S. A. Burikov, S. A. Dolenko, T. A. Dolenko, and I. G. Persiantsev, "The Way to Use the Neural Network Algorithms for Solving the Identification Problems and for Determining Salts Concentration in Multi-Component Water Solutions According to Light Combined Scattering Spectra," Neirokomp'yut.: Razrab. Prim., No. 3, 55–69 (2010).

14. S. A. Burikov, S. A. Dolenko, T. A. Dolenko, and I. G. Persiantsev, "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. Memory Neural Networks (Inform. Opt.) **19** (2), 140–148 (2010).

15. NeuroShell 2, http://www.wardsystems.com/neuroshell2.asp

Sergei Anatol'evich Dolenko. Born 1962. Graduated *cum laude* in 1985 from Physical Department of Lomonosov Moscow State University. Candidate degree in physics and mathematics (2002). Senior research scientist in D.V.Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University. Area of research: elaboration and application of adaptive methods of data analysis (neural networks, genetic algorithms etc.) to solve pattern recognition problems, inverse problems, and other data analysis problems. Author of about 100 articles. Member of Russian Association of Pattern Recognition and Image Analysis, member of Russian Neural Network Society. Member of the editorial board of the journal "Neurocomputers: Development, Applications."

Tat'yana Al'defonsovna Dolenko. Born 1961. Graduated *cum laude* in 1983 from Physical Department of Lomonosov Moscow State University. Candidate degree in physics and mathematics (1987). Senior research scientist in Moscow State University (Department of Physics). Area of research: Raman and IR spectroscopy, aqueous solutions, colloidal solutions of nanoparticles, inverse problems, artificial neural networks. Author of about 85 papers. Member of Russian Association of Pattern Recognition and Image Analysis. Laureate of Lenin Komsomol Prize (1985).

Sergey Alekseevich Burikov. Born 1978. Graduated in 2002 from Physical Department of Lomonosov Moscow State University. Candidate degree in physics and mathematics (2008). Senior research scientist in Lomonosov Moscow State University (Department of Physics). Area of research: Raman and IR spectroscopy, aqueous solutions, colloidal solutions of nanoparticles, inverse problems. Author of 28 papers. Fellow of MSU Scholarship for talented young researchers and lecturers.

Igor' Georgievich Persiantsev. Born 1937. Graduated in 1962 from Physical Department of Lomonosov Moscow State University. Candidate degree in physics and mathematics (1974), doctor degree in physics and mathematics (1988). Leading research scientist in Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University. Area of research: elaboration and application of adaptive methods of data analysis (neural networks, genetic algorithms etc.) to solve pattern recognition problems, prediction problems, and other data analysis problems. Author of about 170 articles. Member of Russian Association of Pattern Recognition and Image Analysis, member of Russian Neural Network Society. Member of the editorial board of this journal, member of the editorial board of the journal "Neurocomputers: Development, Applications." Laureate of the State prize of USSR (1978).