International Journal of Chemical, Materials and Biomolecular Sciences
ISSN: 2415-6620
Vol:8, No:10, 2014

Kohonen Self-Organizing Maps as a New Method for Determination of Salt Composition of Multi-Component Solutions

Sergey A. Burikov, Tatiana A. Dolenko, Kirill A. Gushchin, Sergey A. Dolenko

***Abstract*—**The paper presents the results of clusterization by Kohonen self-organizing maps (SOM) applied for analysis of array of Raman spectra of multi-component solutions of inorganic salts, for determination of types of salts present in the solution. It is demonstrated that use of SOM is a promising method for solution of clusterization and classification problems in spectroscopy of multi-component objects, as attributing a pattern to some cluster may be used for recognition of component composition of the object.

***Keywords*—**Kohonen self-organizing maps, clusterization, multi-component solutions, Raman spectroscopy.

I. Introduction

TODAY the problems of control of aqua technical media; diagnostics and control of water used for irrigation of farmland; control of composition of mineral waters; control of water used for production of beverages; determination of salt composition of sea, river and formation waters, are very urgent [1], [2]. To solve the specified problems, it is necessary to create multipurpose methods and equipment for diagnostics, control and treatment of water, as well as to integrate them into unified technological plans. Obviously, every stage of these technological plans requires sensitive, remote and express methods of determination of salt/ion composition of water.

To obtain operational information about salt/ion composition of water media, the remote methods of diagnostics of water solutions used at present are methods of vibrational spectroscopy - Raman spectroscopy (RS) and spectroscopy of infrared absorption (IR), which allow obtaining information in real-time mode [3]-[5]. At present time, the activity in application of these methods in water media diagnostics is directed towards simultaneous determination of possibly greater number of different ions in water. To solve such multi-parameter inverse problems, adaptive methods of data analysis such as artificial neural networks are successfully used [6], [7].

Authors of [7]-[10] elaborated remote methods of identification of salts and determination of their concentrations in multi-component water solutions by Raman spectra using artificial neural networks (ANN). Use of ANN allowed obtaining high selectivity of the method to the type of dissolved salt/ions and high accuracy of determination of concentration of every salt [9], [10]. The method was elaborated mainly for diagnostics of natural waters – sea, river, mineral waters. In contrast to natural waters, ion composition and concentration in water technological media vary in significantly greater ranges. This means that a method using ANN which was trained for recognition of certain ions may turn to be useless. Such situation necessitates prior determination of salt/ion composition of water media, for example, by clusterization of the data array with consequent correlation of each cluster with certain ion composition. In this paper, Kohonen self-organizing maps (SOM) were used for data clusterization.

Kohonen neural networks and self-organizing maps [11] have been successfully used for a long time for solution of problems of clusterization and visualization of multidimensional data in various areas of human activity [12], [13]. However, till now the authors of this paper found no literature sources describing use of Kohonen networks or SOM to solve the problem of determination of composition of multi-component mixtures. Such studies are carried out for the first time by the authors of this paper.

II. Experimental

To perform elaboration of data processing methods considered in this study, an experimental array of Raman spectra of solutions with various ion compositions (containing from one to five cations and from one to five anions) has been obtained. To facilitate elaboration of the method of solution of the identification problem, solutions with relatively high concentrations of ions were used, in order to maximize as far as possible the distance between spectra of solutions with different ion compositions in the space of intensities in spectral channels.

A. The Objects of Research

The studied object were the aqueous solutions of inorganic salts KI, NH_4F, $NaNO_3$, MgSO4, $AlCl_3$. To prepare solutions,

S. A. Burikov, T. A. Dolenko, and K. A. Gushchin are with the Physical Department, M. V. Lomonosov Moscow State University, Moscow, 119991, Russian Federation (e-mail burikov@lid.phys.msu.ru, tdolenko@lid.phys.msu.ru, kirgush@gmail.com).

S. A. Dolenko is with the D. V. Skobeltsyn Institute of Nuclear Physics, M. V. Lomonosov Moscow State University, Moscow, 119991, Russian Federation (e-mail dolenko@srd.sinp.msu.ru).

This study has been supported by the grant of the Russian Foundation for Basic Research (project no.13-01-00897-a) (S. Burikov, T. Dolenko) and by the grant of the Russian Scientific Foundation (project no.14-11-00579) (S. Dolenko, K. Gushchin). The study has been conducted at the Physical Department, M.V.Lomonosov Moscow State University, and in D. V. Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University, Moscow, 119991, Russian Federation.

bidistilled water and analytically pure reagents were used. The concentration of each salt in the solutions was changed in the range from 0 to 3-4 M (Table I), so that the total concentration would not exceed 4 M. The step of concentration change was 0.05 – 0.1 M for all salts.

TABLE I
CONCENTRATION RANGES OF SALTS IN THE STUDIED SOLUTIONS

Salts	Salt solubility in water, M	Maximum total salts concentration in the solutions, M
KI	6.0	4.0
NH_4F	22.0	4.0
$NaNO_3$	9.0	4.0
$MgSO_4 \cdot 7H_2O$	2.5	2.5
$AlCl_3 \cdot 6H_2O$	3.0	3.0

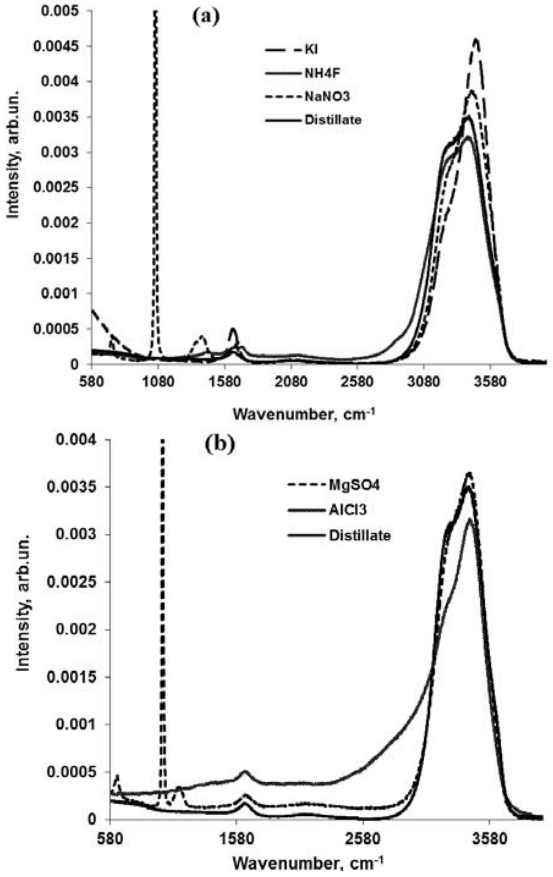

Fig. 1 Raman spectra of water and aqueous single-component solutions of salts with the same concentrations: 3 M - (a), top, and 2 M - (b), bottom

B. Obtaining Raman Spectra of Aqueous Solutions

Raman spectra of aqueous solutions were obtained with the Raman spectrometer described in detail elsewhere [9]. Excitation of Raman spectra was performed by argon laser (wavelength 488 nm, output power 350 mW). In order to remove elastic scattering signal, edge-filter (Semrock) was used. It allowed approaching laser line to 200 cm^{-1}.

Registration of spectra was performed by monochromator (Acton 2500i, grade 900 l/mm, focal length 500 mm) and CCD-camera (Synapse 1024*128 BIUV, Jobin Yvon). Spectra were measured in two regions: 200-2300 cm^{-1} and 2300-4000 cm^{-1} for every sample. Practical resolution of the spectrometer was 2 cm^{-1}, duration of accumulation of one spectrum was 1 s. The temperature of samples during experiment was stabilized at 22.0±0.2°C. Spectra were normalized to laser power, duration of registration of the spectrum, and spectral sensitivity of the detector.

The obtained data array included 807 experimental spectra.

In Figs. 1, 2, examples of obtained Raman spectra of water and aqueous solutions of multicomponent salts are presented. The wide band in the region from 2900 to 4000 cm^{-1} is the band of valence vibrations of water molecules. The band with maximum at 1600 cm^{-1} is produced by bending vibrations of water molecules. As it can be seen from Fig.1, the influence of each ion on the shape and position of Raman spectral bands of water is different. That is a good identification feature for clusterization of the array of spectral data.

Also observed are the vibrational bands of molecular groups of complex ions: SO_4^{2-} (e.g. valence band with maximum at 990 cm^{-1}), NO_3^- (e.g. valence band with maximum at 1060 cm^{-1}), NH_4^+ (bands in the region 2730 cm^{-1}).

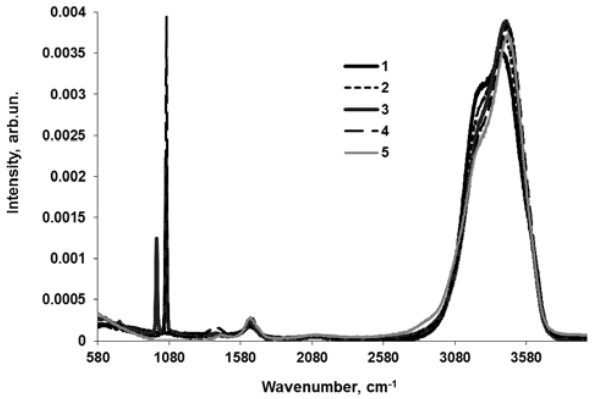

Fig. 2 Raman spectra of water and aqueous multi-component solutions of salts
1: Distilled water; 2: KI - 0.45 M, NH_4F - 0.45 M, $NaNO_3$ - 0.45 M, $MgSO_4$ - 0.35 M, $AlCl_3$ - 0.4 M; 3: KI - 0.65 M, $NaNO_3$ - 0.6 M, $MgSO_4$ - 0.5 M, $AlCl_3$ - 0.6 M; 4: KI - 1.2 M, NH_4F - 1.2 M, $NaNO_3$ - 1.2 M; 5: KI - 1.7 M, NH_4F - 2 M.

It can easily be seen that both the position and the intensity of molecular groups of complex ions may also be identification features for clusterization.

Total number of recorded spectral channels was 1954. In this study, the experiments on clusterization were performed in the space of all 1954 features. However, it is clear from the shape of the spectra that not all the input features are equally informative. Selection or extraction of significant features will be performed in future studies.

III. Use of SOM for Solution of Clusterization Problem

A. Architecture and Parameters of the Neural Networks

In this study, we used the implementation of Kohonen ANN and SOM in the software package Deductor 5.3 – Russian version of Loginom software [14]. The following network parameters were used that turned out to be optimal: the network consisted of 8*8 hexagonal cells with Euclidean distance function; the initial learning rate was 0.5, the finishing learning rate was 0.01; training continued for 500 epochs. In this series of computational experiments, full dimensionality data were used (1954 input features). The number of obtained clusters varied from 2 to 40.

B. Results of Computational Experiments

Fig. 3 presents an example of a SOM obtained for clusterization into 5 clusters. Cell colors density represents average value of the corresponding parameter (concentration of the specified salt, number of salts in the solution etc.) for this cell. Brighter color corresponds to higher parameter value (please refer to the legend bar at the bottom of each map). The density matrix displays the number of patterns that fell into each cell, with no patterns in the two cells covered with dots.

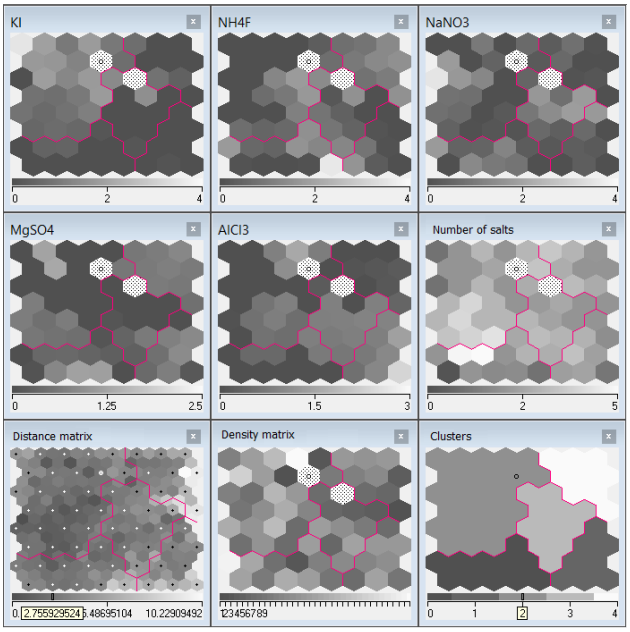

Fig. 3 Sample SOMs for data clusterization into 5 clusters

Fig. 4 shows the change of the fraction of patterns falling into each cluster, depending on the number of clusters.

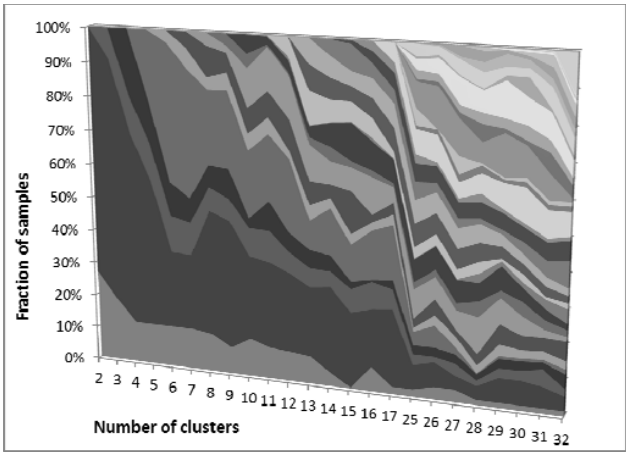

Fig. 4 Fraction of samples falling into each cluster vs number of clusters in the clusterization

To describe quantitative properties of the obtained clusterization, let us introduce the following characteristics.

Let N^i_{max} be the maximum number of patterns from class i that fell into one cluster for the given clusterization, and N the total number of patterns. Let us call the normalized sum over all classes

$$\sum_i N^i_{max} / N,$$

expressed in percent, the *contrast of clusterization*. In the ideal case (if all the samples from each class fall into the same cluster) this index is equal to 100%.

Let us call the ratio of the number of classes, which fell into different clusters by the majority of their samples (i.e. the number of distinguished classes), to the total number of classes (32), expressed in percent, the *degree of separation*.

It is clear that with increasing number of clusters C, the degree of separation grows, tending to 100%, but this growth is non-linear due to existence of unused clusters. On the contrary, contrast tends to decrease.

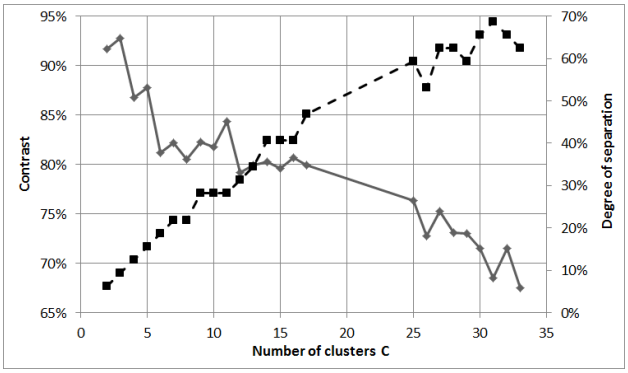

Fig. 5 Contrast (solid line, left axis) and degree of separation (dashed line, right axis) vs number of clusters in the clusterization

Fig. 5 shows the dependences of clusterization contrast (solid line, left axis) and of the degree of separation (dashed

line, right axis) on the number of clusters C ranging from 2 to 33.

Since in this study the investigated solutions contained from 0 to 5 salts, with respect to what combination of salts was present in the solutions belonging to a given class, it was possible to discriminate 2^5=32 data classes. These 32 classes include the following groups: 1 class with spectra of distilled water, 5 classes with spectra of single-salt solutions, 10 classes with two- and 10 classes with three-component solutions spectra, 5 classes with spectra of solutions of 4 salts, 1 class with spectra of solutions containing all the 5 salts. Therefore, one could expect that the most contrast picture would be given by clusterization with the number of clusters close to the following values: 5, 6, 16, 26, 31, 32.

C. Analysis of the Results

The analysis of SOMs and of the results of clusterization obtained with different values of the number of clusters, allow making the following observations and conclusions:

1) Classes corresponding to distilled water, to single-component solutions (solutions of a single salt), and to the solutions of all the five salts, nearly in all experiments fall completely to a single cluster each, to a single or to several adjacent cells of the map. This means that each of these classes form a compact group in the space of input features.
2) The cell corresponding to distilled water, and one or several cells corresponding to five-component solutions, are usually near in the space of the map, and often fall to the same cluster. This may be explained in the following way. If we consider the centroid of the cell with spectra of distilled water to be some kind of an origin of coordinates in the feature space of the problem, increasing concentration of some salt in the solution means movement in this space in some definite direction, along some vector characterizing this salt. If there are several salts in the solutions, this means movement along a vector that is a sum of the characteristic vectors for these salts. If there are many salts, the characteristic vectors may sum up to zero or at least to some vector short enough to be near the origin of coordinates and to fall into the same or adjacent cells with the spectra of distilled water.
3) Locally optimal clusterizations corresponding to local maxima on the contrast curve (Fig. 5) are observed, among other values, near the expected ones – at C = 3, 5, 7, 9, 11, 14, 16, 27, and 32. Note also that the value of contrast keeps at a kind of a plateau in a wide range of C values – from 6 to 16.
4) The formed clusters are non-uniform by their quantitative composition. Thus, samples from the classes corresponding to the solutions containing simultaneously two salts with simple ions (KI and AlCl3), often fall into the same cluster. This may be the manifestation of the fact that changes in the spectrum, introduced by simultaneous presence of these salts, are substantial, and the influence of other salts is less significant against this background.
5) The salt most easily distinguished from all others is KI, which is not surprising from physics and chemical points of view, as this salt has the most pronounced influence on the shape of Raman spectra.
6) This series of computational experiments on data clusterization, conducted on the full array of input features (1954), should be considered as a reference point for future studies, in which similar work will be performed on a reduced number of features, obtained as the result of their selection or extraction. Such method was already applied by the authors before [15] for another data array. It is expected that this will improve the quality of clusterization.
7) More quantitative characteristics should be introduced to describe the clusterization quality in respect to specific salts or their combinations. Unfortunately, calculation of well-know clusterization indexes such as Dunn index, Silhouette, Pearson Gamma index, Entropy and some others [16] gave contradictive results and failed to provide grounds for selection of one or other specific clusterization.

IV. Conclusions

It has been demonstrated that data clusterization with the help of Kohonen neural networks can be used to distinguish groups in the space of intensities in channels of Raman spectra of multi-component solutions of inorganic salts that turn out to be sensible form physical and chemical points of view.

The obtained clusterization may serve as the base for identification of ions (salts) present in the solution. As the borders between adjacent clusters in the feature space are formed in this case in unsupervised learning mode, there are reasons to expect that classification performed on the base of the described clusterization will be more stable than that performed by a neural network trained to distinguish presence of different salts in the solution in supervised learning mode. In future, this hypothesis should be checked on out-of-sample experimental data, which would make it possible to test stability of problem solution against different factors (ion composition of the solution, parameters of the experimental setup and others)/

The main direction of future studies will be search for a feature space of lower dimension, in which clusterization will have optimal properties. Work on elaboration of quantitative indexes of clusterization, informative for the studied problem, should be also continued.

Along with that, further studies will be conducted in the direction of determination of the ion composition of multi-component solution by neural networks with supervised learning.

References

[1] T. R. Crompton, Determination of anions in natural and treated waters. Taylor&Francis, 2002, 828 p.
[2] R. Michalski, "Ion Chromatography as a Reference Method for the Determination of Inorganic Ions in Water and Wastewater," Critical Reviews in Analytical Chemistry, vol. 36, pp. 107–127, 2006.
[3] M. Chaplin, Water Structure and Behavior. www.lsbu.ac.uk/water, 2008

[4] T. A. Gogolinskaia, S. V. Patsaeva, and V. V. Fadeev, "About regularities of change of band 3100-3700 cm-1 of water Raman spectrum of water solutions of salts," DAN USSR (Russian), vol. 290, no. 5, pp. 1099-1103, 1986.
[5] 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," Applied Spectroscopy, vol. 61, no. 12, pp. 274A-292A, Dec. 2007.
[6] A. N. Gorban, V. L. Dunin-Barkovsky et.al., "Neural-network informational models of complex engineering systems" in Neuroinformatica, ch. 4, S.A. Terechov, Ed. Novosibirsk: Nauka. Sibirskoe predpriyatie RAN, 1998.
[7] S. A. Dolenko, T. A. Dolenko, I. G. Persiantsev, V. V. Fadeev, and S. A. Burikov, "Solution of inverse problems of optical spectroscopy using neural networks," Neurocomputers: Development, Application, no. 1-2, pp. 89-97, 2005. (Russian)
[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 Recognition and Image Analysis, vol. 17, no. 4, pp. 554-559, 2007.
[9] S. A. Burikov, S. A. Dolenko et al., "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), vol. 19, no. 2, pp. 140-148, 2010.
[10] S. A. Dolenko, S. A. Burikov, T. A. Dolenko and I.G. Persiantsev, "Adaptive Methods for Solving Inverse Problems in Laser Raman Spectroscopy of Multi-Component Solutions," Pattern Recognition and Image Analysis, vol. 22, no. 4, pp. 551-558, 2012.
[11] T. Kohonen, Self-Organizing Maps. 3d Edition. Berlin: Springer, 2001.
[12] U. Seiffert, and L.C. Jain, Self-Organizing neural networks: recent advances and applications, Heidelberg, New York: Physica-Verlag, 2002.
[13] "Self-Organizing Maps - Applications and Novel Algorithm Design," 2011. http://www.intechopen.com/books/self-organizing-maps-applications-and-novel-algorithm-design
[14] Loginom – analytical platform. http://loginom.basegroup.ru/
[15] S. A. Dolenko, S. A. Burikov, T. A. Dolenko, A. O. Efitorov, and I. G. Persiantsev, "Input data compression at neural-network solution of inverse problems of spectroscopy of multicomponent solutions," in Proc. XV Russian Scientific-Technical Conf. Neuroinformatica (Russian). Moscow, 2013, vol. 2, pp. 205-215.
[16] B. Desgraupes. Clustring Indices. Supplement for Package clusterCrit for R. http://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf