VERSITA

Cent. Eur. J. Chem. • 12(5) • 2014 • 594-603
DOI: 10.2478/s11532-014-0514-6

Central European Journal of **Chemistry**

A procedure for meaningful unsupervised clustering and its application for solvent classification

Research Article

Yaroslava Pushkarova[1], Yuriy Kholin[2*]

[1]*Department of Medical and General Chemistry,*
Bogomolets National Medical University,
01601 Kyiv, Ukraine

[2]*Materials Chemistry Department,*
V. N. Karazin Kharkiv National University,
61022 Kharkiv, Ukraine

Received 29 November 2013; Accepted 5 December 2013

Abstract: Artificial neural networks have proven to be a powerful tool for solving classification problems. Some difficulties still need to be overcome for their successful application to chemical data. The use of supervised neural networks implies the initial distribution of patterns between the pre-determined classes, while attribution of objects to the classes may be uncertain. Unsupervised neural networks are free from this problem, but do not always reveal the real structure of data. Classification algorithms which do not require *a priori* information about the distribution of patterns between the pre-determined classes and provide meaningful results are of special interest. This paper presents an approach based on the combination of Kohonen and probabilistic networks which enables the determination of the number of classes and the reliable classification of objects. This is illustrated for a set of 76 solvents based on nine characteristics. The resulting classification is chemically interpretable. The approach proved to be also applicable in a different field, namely in examining the solubility of C_{60} fullerene. The solvents belonging to the same group demonstrate similar abilities to dissolve C_{60}. This makes it possible to estimate the solubility of fullerenes in solvents for which there are no experimental data

Keywords: *Artificial neural network • Unsupervised classification • Solvents • Fullerene solubility*

1. Introduction

Classification of objects proceeding from their numerical characteristics is considered to be the main tool of modern qualitative chemical analysis [1–3]. Also, classification is widely used to extract useful information from multivariate experimental data for substances, materials, foodstuffs, industrial wastes, environmental objects, etc. [4–7]. Classification methods have become ubiquitous in chemometrics, and novel computational methods and algorithms are continuously being developed. In this context, there are some open fundamental problems rooted in the nature of the numerical analysis of large arrays of raw multidimensional chemical data. The determination of the number of classes in which objects are to be distributed is one of these problems.

In supervised classification, a researcher allocates objects from the training subset between the pre-determined classes, and this assignment is used to build the classification rules. The classes are assumed to be sufficiently well-separated in the data space, and the misclassification of objects from the training subset is excluded [8]. In unsupervised classification, no training data are needed, but the number of classes should be specified [9]. Thus, both supervised and unsupervised classification procedures require information about the number of classes.

In reality, chemists often face the need to analyze novel data sets for which the number of homogeneous classes is unknown and should be determined in the process of data handling. Besides, the rational criteria for the assignment of an object to one or another class may

* *E-mail: kholin@karazin.ua*

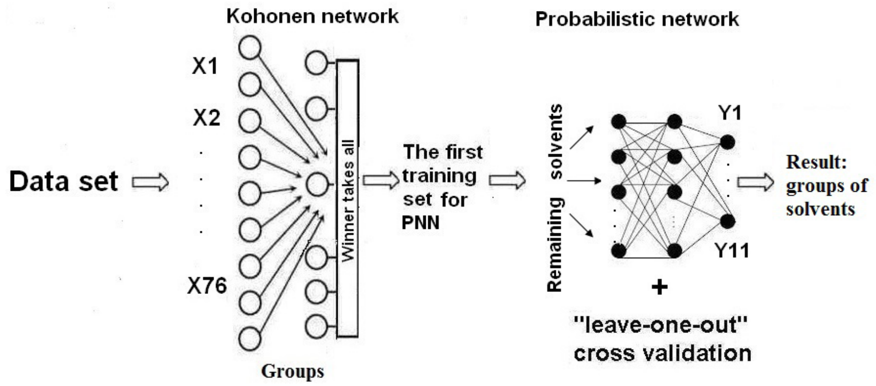

Figure 1. The scheme of the proposed classification procedure based on a combination of the Kohonen and the probabilistic neural networks.

be indistinct. In this case, a combination of chemometric methods is necessary [10–14].

This paper suggests a new approach to the determination of the number of classes based on the combination of the Kohonen and probabilistic artificial neural networks (ANN). Artificial neural networks, due to their well-known advantageous properties, proved to be a powerful classification tool [15–19]. In developing and validating our approach, we addressed the classification problem for solvents and used as many as nine physicochemical characteristics. The classification of solvents is one of the most impressive examples of classification tasks for which an unambiguous answer is impossible. Accordingly, it was important to examine the applicability of the proposed procedure and to obtain meaningful results for this difficult case. Also, the use of nine characteristics of solvents is of special importance, because the commonly used classifications are based on only few parameters [20–22].

2. Theoretical details

2.1. Software and computing

The software package MATLAB 6.5 along with the Neural Networks Toolbox was used in the present work.

2.2. Brief remarks about the Kohonen and probabilistic neural networks

The Kohonen network is a simple two-layer unsupervised network. This type of ANN uses a competitive learning algorithm. During the training stage, the input vector is presented to the network and only one neuron (winning neuron) is activated. The winning neuron is selected as the neuron that has the smallest Euclidean distance to the input vector. Weight coefficients of the winning neuron are modified according to the learning rule [23,24]. A comprehensive description of the competitive learning algorithm was published elsewhere [25]. The Kohonen network is intended for the classification of input vectors into groups; the number of classes must be assigned *a priori* [26].

The probabilistic neural network (PNN) is a type of supervised network that uses radial basis functions for activation (RBFN). RBFNs typically have three layers: the input layer, the hidden layer with a radial basis activation function, and the output layer. Each hidden neuron can store only one sample of the training set [27]. RBFNs are used for solving many problems such as adaptive modeling, approximation, classification and clustering problems [28]. PNN is often used in classification problems with a Gaussian radial basis function as the activation function for its hidden layer. The output layer of PNN is the competitive layer of neurons that determines the most probable class for a given input vector [29].

2.3. Proposed procedure for solving the classification problem

This paper presents an approach to determine the number of classes and to distribute objects between classes. The main idea is to combine the Kohonen and probabilistic neural networks. The Kohonen network is used to determine the number of classes and to form the first training set for PNN. The probabilistic neural network is used to classify the objects which were not included in the first training set. The main steps of the proposed procedure are as follows:

1) to classify objects with the use of the Kohonen neural network at different assigned numbers of classes;

2) to reveal groups of objects which were assigned to the same classes independently of the prescribed number of classes; these objects form the first set for training the PNN;

3) to form testing sets of approximately equal sizes from the remaining objects;

4) to classify objects from each testing set (handled in turn) with the use of PNN; to include classified objects from the testing set into the training set;

5) to verify the obtained classification with the use of a leave-one-out cross-validation procedure.

The proposed classification procedure is depicted in Fig. 1.

2.4. Data set

The values of physicochemical characteristics for 76 organic solvents were taken from the work of Marcus [30]. Solvents are characterized by nine descriptors: solubility, surface tension, dipole moment, relative permittivity, refractive index, hydrogen-bond donation ability, polarity / polarizability, Dimroth-Reichardt polarity index, and structuredness. The values of structuredness for 1-decanol, anisole, hexamethylphosphoramide and propylene carbonate were unknown. These gaps were filled by the average structuredness of the remaining 72 solvents (0.63) (see the values of physicochemical characteristics for 76 organic solvents in the Supplementary Information).

Because of the very wide range of descriptor values, the numerical characteristics of solvents were first autoscaled [31,32]:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}, \; 1 \le i \le 9, \; 1 \le j \le 76,$$

where x_{ij} and $\tilde{x}_{ij}$ are the *i*-th characteristics of the *j*-th solvent before and after the autoscaling; $\bar{x}_i$ and s_i are the mean value and the standard deviation of the i-th characteristic.

3. Results and discussion

3.1. Solvent classification

Results of the classification of 76 solvents with the use of the Kohonen network at different numbers of classes (from *n*=5 to *n*=8) are shown in Table 1.

Ten groups of solvents that were assigned to the same classes independently of the prescribed number of classes were detected (Table 2), and each group consisted of four or more solvents. The total number of such solvents (indicated by italics in Table 1) was 53. These ten groups of solvents were used as the first training set for PNN.

The remaining 23 solvents were randomly subdivided into three testing sets: two sets with 8 solvents and the third set with 7. The PNN was used to handle the testing sets sequentially, *i.e.*, solvents from the first testing set were transferred to a certain training set, after which the classification of objects from the second testing set was performed, and finally solvents from the third set were classified. As such, the size of training set increased at each step of the procedure.

Classification of solvents from the three testing sets is presented in Tables 3–5.

3.2. Verification of the resulting classification

The leave-one-out cross-validation method was used in this work [33]. The original data set was randomly subdivided into 5 sub-sets: four sub-sets with 15 objects and one with 16 objects. One sub-set was retained as the testing set, and the remaining 4 sub-sets were used as the training sets. The cross-validation process was repeated 5 times in such a way that each sub-set was used once as the testing set.

As a result of the leave-one-out cross-validation, 12 solvents were transferred from one group to another (Table 6).

Note that chloroform and 1-aminobutane were assigned to Group 1, which consists of apolar aliphatic hydrocarbons, in spite of obvious dissimilarity of their properties. Most likely, this happens because there were no appropriate groups for these two solvents in the training set. To check this, an additional group has been created and verified. Chloroform was transferred to the training set as a member of this group, and 1-aminobutane was used as the testing object. As a result of the PNN application, 1-aminobutane was assigned to the group of chloroform. The assignment of chloroform to a group different from other halogen-substituted hydrocarbons agrees with Snyder and Chastrette's solvent classification schemes [34,35].

3.3. Discussion of the resulting classification of solvents

The final classification resulting from the joint use of the Kohonen neural network, probabilistic neural network~~s~~, and the leave-one-out cross-validation is shown in Table 7. We discuss some physico-chemical arguments that substantiate it.

Apolar and slightly polar solvents were divided into Groups I-III. Group I consists of aprotic apolar aliphatic hydrocarbons with less than 10 carbon atoms, and hexafluorobenzene. Group II includes apolar and slightly polar aliphatic hydrocarbons more than 10 carbon atoms, aromatic and cyclic hydrocarbons, tetrachloromethane and trichloroethylene. The aliphatic hydrocarbons mentioned have higher relative permittivities and refractive indices than hydrocarbons from Group I. This explains splitting of the aliphatic

Table 1. Classification of solvents with the use of the Kohonen network.

№	Solvent	Class			
		***n*=5**	***n*=6**	***n*=7**	***n*=8**
1	*n-pentane*	*1*	*4*	*2*	*1*
2	*n-hexane*	*1*	*4*	*2*	*1*
3	*n-heptane*	*1*	*4*	*2*	*1*
4	*n-octane*	*1*	*4*	*2*	*1*
5	*i-octane*	*1*	*4*	*2*	*1*
6	*n-decane*	*1*	*4*	*2*	*1*
7	*n-dodecane*	*1*	*4*	*2*	*1*
8	n-tetradecane	1	4	2	2
9	n-hexadecane	1	4	2	2
10	*c-hexane*	*1*	*4*	*2*	*1*
11	*methyl-c-hexane*	*1*	*4*	*2*	*1*
12	cis-decalin	1	5	4	2
13	*benzene*	*3*	*5*	*4*	*2*
14	*p-xylene*	*3*	*5*	*4*	*2*
15	*mesitylene*	*3*	*5*	*4*	*2*
16	hexafluorobenzene	1	4	1	2
17	tetrachloromethane	1	5	2	2
18	*trichloroethylene*	*3*	*5*	*2*	*2*
19	*tetrachloroethylene*	*3*	*5*	*2*	*2*
20	*toluene*	*3*	*5*	*2*	*2*
21	*o-xylene*	*3*	*5*	*2*	*2*
22	*m-xylene*	*3*	*5*	*2*	*2*
23	ethylbenzene	3	5	1	2
24	*cumene*	*3*	*5*	*4*	*2*
25	*1,4-dioxane*	*3*	*1*	*1*	*6*
26	fluorobenzene	3	1	1	8
27	dichloromethane	2	1	1	3
28	chloroform	3	1	1	3
29	1,2-dichloethane	2	1	5	8
30	*1,1,1-trichloroethane*	*3*	*1*	*1*	*6*
31	1,1,2,2-tetrachloroethane	4	2	5	8
32	1-chloropropane	1	1	1	3
33	*1,2,3-trichloropropane*	*4*	*1*	*5*	*8*
34	*chlorobenzene*	*4*	*1*	*5*	*8*
35	o-dichlorobenzene	4	2	5	5
36	m-dichlorobenzene	4	5	7	8
37	*1,2,4-trichlorobenzene*	*4*	*2*	*7*	*5*
38	*dibromomethane*	*4*	*2*	*7*	*5*
39	*bromoform*	*4*	*2*	*7*	*5*
40	*1,2-dibromoethane*	*4*	*2*	*7*	*5*
41	*1-bromopropane*	*3*	*1*	*1*	*6*

Continued **Table 1.** Classification of solvents with the use of the Kohonen network.

№	Solvent	Class			
		n=5	*n*=6	*n*=7	*n*=8
42	*bromobenzene*	4	2	7	5
43	*diiodomethane*	4	2	7	5
44	*1-iodopropane*	4	1	5	8
45	*iodobenzene*	4	2	7	5
46	*tetrahydrofuran*	3	1	1	6
47	*anisole*	4	1	5	8
48	1-aminobutane	1	1	1	3
49	pyridine	2	2	5	8
50	*quinoline*	4	2	7	5
51	carbon disulfide	4	5	7	5
52	*tetrahydrothiophene*	4	1	5	8
53	*methanol*	5	6	3	7
54	*ethanol*	5	6	3	7
55	*1-propanol*	5	6	3	7
56	*1-butanol*	5	6	3	3
57	*1-pentanol*	5	6	3	3
58	*1-hexanol*	5	6	3	3
59	*1-octanol*	5	6	3	3
60	*1-decanol*	5	6	3	3
61	1,2-ethanediol	5	6	6	7
62	water	5	3	6	7
63	N-methylformamide	5	3	6	7
64	*N,N-dimethylformamide*	2	3	6	4
65	*N,N-dimethylacetamide*	2	3	6	4
66	*N-methylpyrrolidone*	2	3	5	4
67	*hexamethylphosphoramide*	2	3	5	4
68	*dimethylsulfoxide*	2	3	6	4
69	*o-cresol*	5	6	3	7
70	*propylene carbonate*	2	3	6	4
71	acetone	2	6	3	3
72	nitromethane	5	3	6	7
73	*nitroethane*	2	3	6	4
74	*nitrobenzene*	2	3	5	4
75	acetonitrile	5	6	6	7
76	*benzonitrile*	2	3	5	4

hydrocarbons into two groups, and inclusion of aliphatic hydrocarbons with a higher number of carbon atoms in the same group containing aromatic and cyclic hydrocarbons. Hexafluorobenzene was assigned to Group 1 due to the values of its relative permittivity and refractive index.

Group III consists of only cis-decalin, carbon disulfide, and tetrachloroethylene. Cis-decalin has a higher surface tension than cyclic hydrocarbons. Two polyhalogen-substituted aliphatic hydrocarbons (tetrachloromethane and tetrachloroethylene) were assigned to different groups due to their surface tension values.

Table 2. The first training set for PNN.

Group	Solvent (number of solvent in Table 1)
I	n-pentane (1), n-hexane (2), n-heptane (3), n-octane (4), i-octane (5), n-decane (6), n-dodecane (7), c-hexane (10), methyl-c-hexane (11)
II	benzene (13), p-xylene (14), mesitylene (15), cumene (24)
III	trichloroethylene (18), tetrachloroethylene (19), toluene (20), o-xylene (21), m-xylene (22)
IV	1,4-dioxane (25), 1,1,1-trichloroethane (30), 1-bromopropane (41), tetrahydrofuran (46)
V	1,2,3-trichloropropane (33), chlorobenzene (34), 1-iodopropane (44), anisole (47), tetrahydrothiophene (52)
VI	1,2,4-trichlorobenzene (37), dibromomethane (38), bromoform (39), 1,2-dibromoethane (40), bromobenzene (42), diiodomethane (43), iodobenzene (45), quinoline (50)
VII	methanol (53), ethanol (54), 1-propanol (55), o-cresol (69)
VIII	1-butanol (56), 1-pentanol (57), 1-hexanol (58), 1-octanol (59), 1-decanol (60)
IX	N,N-dimethylformamide (64), N,N-dimethylacetamide (65), dimethylsulfoxide (68), propylene carbonate (70), nitroethane (73)
X	N-methylpyrrolidone (66), hexamethylphosphoramide (67), nitrobenzene (74), benzonitrile (76)

Table 3. Classification of solvents from the first testing set.

Group	Solvent (number of solvent in Table 1)
II	n-tetradecane (8)
III	cis-decalin (12), ethylbenzene (23), carbon disulfide (53)
IV	1,2-dichloethane (29)
VII	acetone (71)
IX	nitromethane (72)

Table 4. Classification of solvents from the second testing set.

Group	Solvent (number of solvent in Table 1)
I	hexafluorobenzene (16)
II	n-hexadecane (9), tetrachloromethane (17)
V	1,1,2,2-tetrachloroethane (31)
VIII	1-chloropropane (32)
IX	acetonitrile (75), 1,2-ethanediol (61), water (62)

Table 5. Classification of solvents from the third testing set.

Group	Solvent (number of solvent in Table 1)
I	chloroform (28), 1-aminobutane (48)
IV	fluorobenzene (26), dichloromethane (27)
V	o-dichlorobenzene (35)
VI	m-dichlorobenzene (36), pyridine (49)
IX	N-methylformamide (63)

Weakly polar solvents form Groups IV-VII. Groups IV-VI consist of weakly polar halogen-substituted aliphatic and aromatic hydrocarbons, heterocycles (quinoline, pyridine), aromatic and cyclic ethers (tetrahydrofuran, anisole, 1,4-dioxane), tetrahydrothiophene, an o-cresol. These classes of organic substances fall into three different groups due to the specific combinations of their properties. But it is noteworthy that values of the refractive index increase from Group IV to Group VI. Trichloroethylene differs from halogen-substituted aliphatic hydrocarbons of Groups IV-VI by lower values of the dipole moment, relative permittivity and structuredness. It seems that trichloroethylene was assigned to Group II for this reason. o-Cresol is the only representative of phenols. It differs from aliphatic alcohols (Groups VII, VIII) by higher values of surface tension, refractive index, hydrogen-bond-donation ability, polarity / polarizability, and structuredness, and has a lower dipole moment.

Groups VII and VIII contain hydrogen bond donors, acetone and 1-chloropropane. Group VII consists of alcohols with less than 5 carbon atoms and acetone. Group VIII includes alcohols with 5 to 10 carbon atoms and 1-chloropropane. The latter is characterized by the lowest values of surface tension and refractive index among all classified halogen-substituted aliphatic and aromatic hydrocarbons. This justifies the presence of 1-chloropropane in Group VIII.

Groups IX and X are composed of highly polar solvents. Solvents from these groups differ significantly in relative permittivities, Dimroth-Reichardt polarity indices, and refractive indices.

Chloroform and 1-aminobutane form Group XI. Chloroform has the highest hydrogen-bond-donation ability among all halogen-substituted aliphatic and

Table 6. Results of the leave-one-out cross-validation.

Solvent (number of solvent in Table 1)	Group	
	before cross-validation	after cross-validation
m-xylene (22), toluene (20), o-xylene (21), trichloroethylene (18), ethylbenzene (23)	3	2
o-cresol (69)	7	4
c-hexane (10), n-dodecane (7), methyl-c-hexane (11)	1	2
pyridine (49)	6	5
anisole (47)	5	6
1-butanol (56)	8	7

Table 7. The final classification of solvents. Values in the parenthesis are logarithms of C_{60} mole fractions (*x*) in the saturated solutions at 298 K; if the source of data is absent, the *x* values were taken from [42]*.

Group	Solvent	
I	apolar and slightly polar	n-pentane (-6.1), n-hexane (-5.1), n-heptane (?), n-octane (-5.2), i-octane (-5.2), n-decane (-4.7), hexafluorobenzene (?)
II		n-tetradecane (-4.3), n-hexadecane (?), n-dodecane (-3.5), c-hexane (-5.3), methyl-c-hexane (-4.5), trichloroethylene (-3.8), tetrachloromethane (?), benzene (-4.0), p-xylene (-3.3), mesitylene (-3.5), cumene (-3.6), toluene (-3.4), o-xylene (-2.9), m-xylene (-3.3), ethylbenzene (-3.4)
III		cis-decalin (-3.3 [47]), carbon disulfide (-3.2 [47]), tetrachloroethylene (-3.8)
IV	weakly polar	fluorobenzene (-4.1 [47]), dichloromethane (-4.6), 1,2-dichloroethane (-5.0), 1,1,1-trichloroethane (-4.7), 1-bromopropane (-5.2), tetrahydrofuran (?), **o-cresol (-5.7)**, 1,4-dioxane (-5,3 [46])
V		1,1,2,2-tetrachloroethane (-3.1), 1,2,3-trichloropropane (-4.0), chlorobenzene (-3.0), **o**-dichlorobenzene (-2.4), 1-iodopropane (-4.6), pyridine (-4.0), tetrahydrothiophene (-5.4)
VI		bromobenzene (-3.3), m-dichlorobenzene (-3.4), bromoform (-3.2), iodobenzene (-3.5), quinoline (-2.9), 1,2,4-trichlorobenzene (-2.8), **diiodomethane (-4.8),** 1,2-dibromoethane (-4.2), dibromomethane (-4.5), ainnisole (-3.1)
VII	hydrogen bond donors and others	methanol (practically insoluble [39]), ethanol (-7.1), 1-propanol (-6.4), 1-butanol (-5.9), acetone (-7.0)
VIII		1-pentanol (-5.3), 1-hexanol (-5.1), 1-octanol (-5.0), 1-decanol (?), 1-chloropropane (-5.6)
IX	highly polar	N-methylpyrrolidinone (-3.9 [47]), hexamethylphosphoramide (?), nitrobenzene (-3.9), benzonitrile (-4.2)
X		1,2-ethanediol (?), water (practically insoluble [39]), N-methylformamide (?), N,N-dimethylformamide (-5.3), nitroethane (-6.7), N,N-dimethylacetamide (?), dimethylsulfoxide (?), propylene carbonate (?), nitromethane (practically insoluble [39]), acetonitrile (practically insoluble [39])
XI	miscellaneous	chloroform (-4.8), **1-aminobutane (-3.3)**

* *The solvents which are considered as outliers [42] are boldfaced.*

aromatic hydrocarbons considered in this work. At the same time, some characteristics of 1-aminobutan and chloroform (for example, refractive index and relative permittivity) are rather close.

The resulting classification of solvents is similar to Gramatica's [11], Chastrette's [35] and Katritzky's [36,37] solvent classification schemes built for other sets of solvents and their descriptors: trichloroethylene and other halogen-substituted hydrocarbons are placed into different groups, which is in agreement with Chastrette's classification; the composition of Groups VII-X corresponds to Gramatica's, Chastrette's and Katritzky's classifications; as in Katritzky's [36] classification, ethers and halogen-substituted hydrocarbons were placed in the same group; the set of substances in Group III corresponds to Gramatica's classification. In agreement with Gramatica's classification, tetrachloromethane, 1-chloropropane, and other halogen-substituted hydrocarbons were distributed between different groups. Aliphatic and aromatic hydrocarbons also form different groups, similarly to Chastrette's and Katritzky's solvent classifications, our classification also contains Group XI of substances with an uncertain function ("miscellaneous solvents").

To verify the reliability of the suggested approach, it is necessary to test it with another set of experimental data. The descriptors considered in this work include physicochemical characteristics which are used to discuss solubility phenomena and solute-solvent interactions [21,38]. The problem of fullerenes solubility in different solvents is attracting growing attention (see, for instance, recent comprehensive reviews [39,40]). At present, experimental data about the solubility of C_{60} in approximately 150 neat solvents at room temperature are available [39]. Their interpretation is complicated by the considerable disagreements in the reported values [39,40]. Besides, the dissolution of C_{60} in polar solvents is assumed to result in colloid rather than true solutions [40], and in some cases chemical interactions between solvents and C_{60} have been reported [41,42]. As a result, it is difficult to state unambiguously how the solubility of fullerenes depends on solvent parameters; the predictive QSPR models for fullerene solubility are very different, whereas models based on molecular descriptors are rather sophisticated and involve at least four or five quantum-chemical and topological parameters [42-45].

With this in mind, it is reasonable to ascertain whether the assignment of a solvent to a particular group corresponds to a certain level of C_{60} solubility. The initial conjecture is that solvents belonging to the same group have similar properties, including their ability to dissolve C_{60}. The solubility values for C_{60} expressed as its molar fractions (x) in saturated solutions at 298 K are given in Table 7. This dataset was originally presented by Beck and Mandy [46] and subsequently used by many authors [41,43-45].

Proceeding from these data, fullerene C_{60} is *very slightly soluble* in solvents assigned to Group VII of our classification (log x ≤ -5.9), Group VIII (-5.6 ≤ log x ≤ -5.0), Group 1 (-6.1 ≤ log x ≤ -4.7), and Group X (log x ≈ -5 − -7). It is *slightly soluble* in solvents belonging to Group IV (-5.3 ≤ log x ≤ -4.1, o-cresol was excluded from the consideration as the outlier solvent [42]) and *sparingly soluble* in solvents which constitute Groups II, III, VI and IX (-4.6 ≤ log x ≤ -2.8, c-hexane and tetrahydrothiophene demonstrate anomalous dissolving power; diiodomethane was excluded as the outlier solvent [42]).

Thus, from 58 solvents for which experimental information about fullerene solubility is available, the dissolving power of only two solvents does not fit the suggested classification. On the one hand, this confirms the validity of the classification. On the other hand, this makes it possible to estimate semiquantitatively the solubility of fullerenes in solvents for which there are no experimental data. For instance, one can expect that the molar fractions of C_{60} in the saturated solutions in 1,2-ethanediol, N-methylformamide, N,N-dimethylacetamide or propylene carbonate (Group X) do not exceed 10^{-5}.

4. Conclusions

The approach based on the combination of the Kohonen neural network and probabilistic neural network has been studied for its ability to determine rationally the number of classes and to establish the reliable classification of objects in complex classifications problems. The proposed procedure combines advantageous features of supervised and unsupervised classification methods. It does not require any *a priori* information about the number of classes or patterns in the training set. Also, it can handle incomplete data (datasets with gaps) and can treat multiparametric data without compression. The procedure can be considered a prospective tool for the solution of ill-posed chemical classification tasks.

The approach has been demonstrated to be efficient for the classification of a large set of solvents. The additional use of the leave-one-out cross-validation procedure has improved the results. The final solvent classification is meaningful and chemically interpretable. It has been extended to a different problem, the solubility of fullerene C_{60} in neat solvents at 298 K. This provided additional arguments in favor of the validity and efficiency of the suggested approach. It has been demonstrated that solvents belonging to the same group have similar abilities to dissolve C_{60}. This makes it possible to estimate semiquantitatively the solubility of fullerenes in solvents for which there are no experimental data.

Acknowledgements

This work was partly supported by the Ministry of Education and Science of Ukraine (grant 0112U008337). The authors wish to express their deep gratitude to Professor A. Korobov for manuscript revision and to Professor N. Mchedlov-Petrossyan for helpful advice.

References

[1] Yu. Vlasov, A. Legin, A. Rudnitskaya, C. Di Natale, A. D'Amico, Pure Appl. Chem. 77, 1965 (2005)

[2] W.A. Hardcastle, Qualitative analysis: a guide to best practice (Royal Society of Chemistry, Cambridge, 1998)

[3] B.L. Milman, Trends Anal. Chem. 24, 493 (2005)

[4] M.J. Adams, Chemometrics in analytical spectroscopy, 2nd edition (Royal Society of Chemistry, Cambridge, 2004)

[5] Ph.K. Hopke, Anal. Chim. Acta. 500, 365 (2003)

[6] E.F. Boffo, L.A. Tavares, A.C.T. Tobias, M.M.C. Ferreira, A.G. Ferreira, Food Sci. Technol. 49, 55 (2012)

[7] S.S. Souza, A.G. Cruz, E.H.M. Walter, J.A.F. Faria, R.M.S. Celeghini, M.M.C. Ferreira, D. Granato, A. de S. Sant'Ana, Food Chem. 124, 692 (2011)

[8] P.K. Shivaswamy, T. Jebara, J. Mach. Learn. Res. 11, 747 (2010)

[9] R. Wehrens, Chemometrics with R: multivariate data analysis in the natural sciences and life sciences (Springer, Heidelberg, Dordrecht, London, New York, 2011)

[10] A. de Juan, G. Fonrodona, E. Casassas, Trends Anal. Chem. 16, 52 (1997)

[11] P. Gramatica, N. Navas, R. Todeschini, Trends Anal. Chem. 18, 461 (1999)

[12] V. Simeonov, P. Simeonova, S. Tsakovskii, V. Lovchinov, J. Water Resource Protect. 2, 353 (2010)

[13] R. Skorek, M. Jablonska, M. Polowniak, A. Kita, P. Janoska, F. Buhl, Cent. Eur. J. Chem. 10, 71 (2010)

[14] J.S. dos Santos, N.S. dos Santos, M.L.P. dos Santos, S.N. dos Santos, J.J. de Jesus Lacerda, J. Braz. Chem. Soc. 19, 502 (2008)

[15] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Anal. Chim. Acta 671, 27 (2010)

[16] R.M. Alonso-Salces, C. Herrero, A. Barranco, L.A. Berrueta, B. Gallo, F. Vicente, Food Chem. 93, 113 (2005)

[17] P.H. Fidencio, I. Ruisanchez, R.J. Poppi, Analyst 126, 2194 (2001)

[18] R.M. Balabin, R.Z. Safieva, Fuel. 87, 2745 (2008)

[19] O.F. Galao, D. Borsato, J.P. Pinto, J.V. Visentainer, M.C. Carrao-Panizzi, J. Braz. Chem. Soc. 22, 142 (2011)

[20] Y. Marcus, Chem. Soc. Rev. 22, 409 (1993)

[21] Ch. Reichardt, Solvents and solvent effects in organic chemistry, 3rd edition (Wiley & Co. KGaA, Weinhem, 2003)

[22] V.J. Barwick, Trends Anal. Chem. 16, 293 (1998)

[23] F. Marini, J. Zupan, A.L. Magri, Anal. Chim. Acta 544, 306 (2005)

[24] J. Zupan, M. Novic, I. Ruisanchez, Chemometr. Intell. Lab. 38, 1 (1997)

[25] C. Fyfe, Artificial neural networks (Department of Computing and Information Systems, University of Paisley, Scotland, 1996)

[26] W.J. Melssen, J.R.M. Smits, L.M.C. Buydens, G. Kateman, Chemometr. Intell. Lab. 23, 267 (1994)

[27] D. Anderson, G. McNeill, Artificial neural networks technology (Kaman Sciences Corporation, New York, 1992)

[28] M.M. Gupta, L. Jin, N. Homma, Static and dynamic neural networks: from fundamentals to advanced theory (Wiley & Sons, New Jersey, 2003)

[29] W. Tylmon, G.J. Anders, Energy 31, 2874 (2006)

[30] Y. Marcus, J. Phys. Chem. B 101, 8617 (1997)

[31] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, BMC Genomics 7, 142 (2006)

[32] K.R. Beebe, J.P. Randy, M.B. Seasholtz, Chemometrics: a practical guide (Wiley & Sons, USA, 1998)

[33] M. Dong, N. Wang, Appl. Math. Model. 35, 1024 (2011)

[34] V.J. Barwick, Trends Anal. Chem. 16, 293 (1997)

[35] M. Chastrette, M. Rajzmann, M. Chanon, K.F. Purcell, J. Am. Chem. Soc. 107, 1 (1985)

[36] A.R. Katritzky, T. Tamm, Y. Wang, M. Karelson, J. Chem. Inf. Comput. Sci. 39, 692 (1999)

[37] A.R. Katritzky, D.C. Fara, M. Kuanar, E. Hur, M. Karelson, J. Phys. Chem. A 109, 10323 (2005)

[38] Y. Marcus, The properties of solvents. Wiley Series in Solution Chemistry (Wiley, Chichester, 1998) Vol. 4

[39] K.N. Semenov, N.A. Charykov, V.A. Keskinov, A.K. Piartman, A.A. Blokhin, A.A. Kopyrin, J. Chem. Eng. Data 55, 13 (2010)

[40] N.O. Mchedlov-Petrossyan, Chem. Rev. 113, 5149 (2013)

[41] I.Z. Kiss, G. Mandi, M.T. Beck, J. Phys. Chem. A 104, 8081 (2000)

[42] H. Liu, X. Yao, R. Zhang, M. Liu, Z. Hu, B. Fan, J. Phys. Chem. B 109, 20565 (2005)

[43] F. Gharagheizi, R.F. Alamdari, Fullerenes, Nanotubes, and Carbon Nanostructures 16, 40 (2008)

[44] T. Petrova, B.F. Rasulev, A.A. Toropov, D. Leszczynska, J. Leszczynski, J. Nanopart. Res. 13, 3235 (2011)

[45] S.Yousefinejad, F. Honarasa, F. Abbasitabar,

Z. Arianezhad, J. Solution Chem. 42, 1620 (2013)

[46] M.T. Beck, G. Mandi, Fullerenes, Nanotubes, and Carbon Nanostructures 5, 291 (1997)

[47] R.S. Ruoff, D.S. Tse, R. Malhotra, D.C. Lorents, J. Phys. Chem. 97, 3379 (1993)