ELSEVIER

Journal of Chromatography A, 925 (2001) 31–40

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Application of self-organizing maps for the classification of chromatographic systems and prediction of values of chromatographic quantities

Željko Debeljak[a], Martina Strapač[b], Marica Medić-Šarić[c,*]

[a]*Analytical Development Department, PLIVA d.d., Prilaz baruna Filipovića 25, 10 000 Zagreb, Croatia*
[b]*Quality Assurance Department, PLIVA d.d., Prilaz baruna Filipovića 25, 10 000 Zagreb, Croatia*
[c]*Department of Pharmaceutical Chemistry, Faculty of Pharmacy and Biochemistry, University of Zagreb, A. Kovačića 1, 10 000 Zagreb, Croatia*

Received 19 February 2001; received in revised form 30 May 2001; accepted 31 May 2001

Abstract

The applicability of self-organizing maps (SOM) for the classification of chromatographic systems or components of chromatographic systems based on data taken from literature is shown. The SOM approach is compared to dendrogram and principal components analysis (PCA) approaches. It has been shown that the distance between classified objects could reveal linear correspondence with quantity to be optimized, e.g. resolution, so it can be applied in the chromatographic method development. SOMs can also be applied for prediction of chromatographic quantities. It is shown that SOM-based response surface modeling is comparable to triangular presentation of mobile phase composition response surfaces.

Keywords: Self-organizing maps; Principal component analysis; Response surface modeling

1. Introduction

Classification is a widely used tool in chromatography, especially in the early phases of method development [1–7]. Though many people use classification and 'similarity' in an intuitive way there are few theoretically well-founded similarity measures and classification tools like dendrograms [2,6] or PCA [3,7] that have found widespread use. Yet, several questions are to be addressed about the applicability of these tools regarding the chromatographic quantities to be modified. In other words, the limitations of a particular classification and/or classification tool against chromatographic properties should be stated. It is hard to believe that any chromatographic system classification can be universally applicable over the wide range of analytes and for the optimization of large number of different chromatographic quantities. Some insight into the applicability of particular chromatographic system classification gives certain measures of classification quality like the percent of variance that is explained by, for example, the first two principal components. The existence of such measures makes a difference between different classification tools.

In the last decade many different machine-based

*Corresponding author. Tel.: +385-1-481-8304; fax: +385-1-485-6201.

E-mail address: bebamms@nana.pharma.hr (M. Medić-Šarić).

0021-9673/01/$ – see front matter
PII: S0021-9673(01)01010-X

learning tools have been developed, especially different paradigms of supervised learning neural nets, whose applicability has been proved for regression and prediction in many areas including chromatography [8–10]. While those tools can be applied for classification tasks, the unsupervised learning-based approach seems to be more appealing for this purpose. It is very important to make a classification that is as robust as possible against changes of chromatographic properties one wants to modify or optimize based on a particular classification of chromatographic systems or chromatographic components classification. That task is just the opposite to the usual tasks of supervised learning. The cost is more or less lowered prediction accuracy compared to the supervised learning techniques depending on the property one is trying to modify and input data. Of course, one would expect that the solution to that task should be taken with greater caution, and that is the reason why classification is better suited for early phases of method development than for later phases. It simply gives a good starting point for method development because of its robustness against different quantities one wants to optimize while it can be followed by supervised learning techniques in later phases of method optimization.

A number of unsupervised learning techniques have been used in different areas of science but the most usual one is SOM, which has also been applied in different areas of chemistry like molecular modeling and drug design [11,12], identification of compounds based on mass spectra [11], etc. Quite surprisingly, we have managed to find only one paper on the application of SOMs in chromatography [1]. Since SOM is a member of unsupervised learning techniques it suits previous pointed chromatographic systems classification tasks, and it also satisfies the need for a classification quality measure. Since dendrograms and PCA became benchmarks of classification algorithms it is appealing to compare the possibilities of SOM against the possibilities of these tools. The nice thing about SOM is that it can easily be converted to a prediction tool [12] and used for response surface modeling and chromatographic property optimization. So, based on the same data one can apply a classification of chromatographic systems and prediction of experimental outcome at the same time.

2. Theory

A graphical representation of SOM training is given in Fig. 1. It can be seen that every *i*th object that is supposed to be classified is represented, in our case with a three-dimensional vector ($\mathbf{x}_i$). Of course the dimension ($j=1, 2, \ldots, m$) of the vector whose components are in our case chromatographic quantities, is variable and its value is defined by the analyst based on some previous assumptions that we will deal with later in the text. Neurons are represented by circles that form an array with 3×3 dimension. The array, called the feature map does not need to be quadratic, as is the case in Fig. 1. Arrows pointing to every neuron suggest comparison of $\mathbf{x}_i$ with a vector that represents the *k*th neuron ($\mathbf{w}_k$) and the outcome of that comparison is represented by arrows directed outwards. It is obvious that the dimensions of those two kinds of vectors have to be the same. The 'comparison' can be mathematically described with some distance measure $\Delta_{i,k}$:

$$\Delta_{i,k} = |\mathbf{x}_i - \mathbf{w}_k| \tag{1}$$

Since the map training is an iterative process, it is necessary to initialize $\mathbf{w}_k$ values at the beginning of the training. The most important fact about the map training is the 'winner takes the most' principle:

$$i(\mathbf{x}_i) = \arg \min_k \Delta_{i,k} \tag{2}$$

This heuristic states that the neuron that is the closest to the specific $\mathbf{x}_i$ is the winner $i(\mathbf{x}_i)$ in a sense that it changes its vector components in the most significant manner as a result of the *n*th iteration step. But neurons in some predefined neighborhood of the winner also change their vector components and both types of change are described by Eq. (3):

$$\mathbf{w}_k(n+1) = \mathbf{w}_k(n) + \eta(n) h_{k,i(\mathbf{x}_i)}(n) \Delta_{k,i(\mathbf{x}_i)}(n) \tag{3}$$

In this equation we can see two new quantities, each predefined by the analyst. η is known as the learning-rate parameter. It defines how the rate of change of $\mathbf{w}_i$ depends on the value of *n*. The learning-rate function usually has an exponential form. $h_{k,i(\mathbf{x}_i)}$ stands for the amount of change of $\mathbf{w}_k$ in the neighborhood of the $i(\mathbf{x}_i)$ and it is a function of iteration step, interneuron distance and the feature

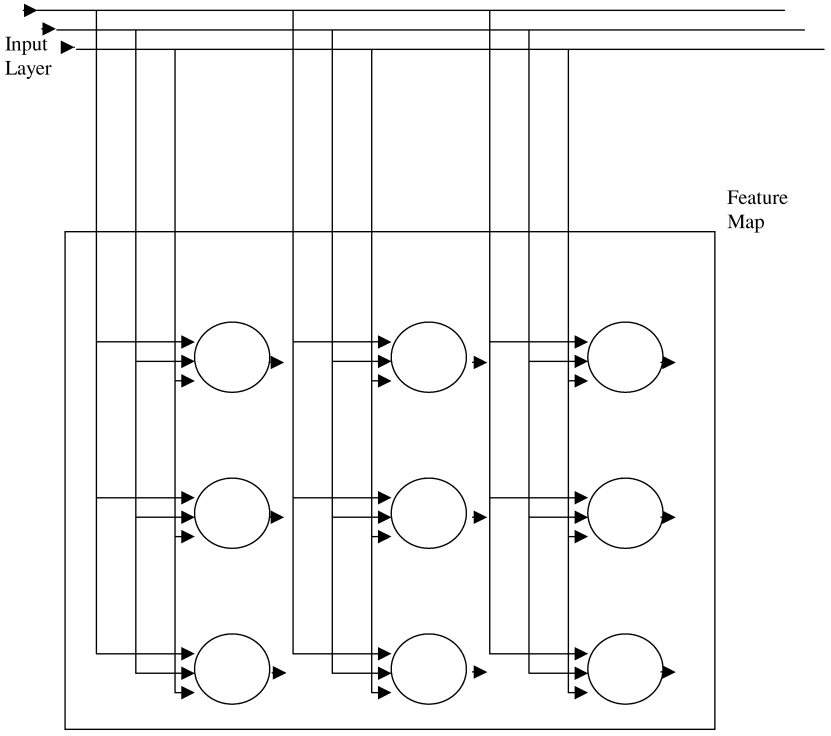

Fig. 1. Graphical presentation of the SOM training algorithm.

map topology, e.g. the way of connecting neurons. The most usual map topologies are rectangular, hexagonal or toroidal. More about learning rate functions, network topologies and other related matters can be found in Refs. [13–15].

Though it is an unsupervised learning technique, SOM can also be modified for prediction [12]. This is based on the fact that the normalized difference $\Delta_{i,k}$ can easily be transformed to a probability measure of finding the *i*th object located at the *k*th neuron. Multiplication of this probability function by the chromatographic quantity that one wants to optimize followed by summation over all objects gives the prediction of the quantity for all objects located at the specific neuron.

3. Experimental

Data chosen for testing of SOM applicability in chromatography method development was taken from literature [2,5,16–19]. Three HPLC and three TLC datasets were used to analyze the general applicability of this approach. Depending on the specific dataset and the purpose of the analysis, we have taken capacity factors/retention times or mobile phase composition as input vector $\mathbf{x}_i$ components. Only the dataset taken from Ref. [5] contains input variables from different origins, i.e. seven physical chemistry descriptors of C_8 and C_{18} HPLC columns. For testing of applicability of SOM in method development and for comparison with dendrogram classification of chromatographic objects we used discriminating power (DP_i) with 5% error factor as a measure of resolution [2]. For comparison of SOM-based prediction with multiple linear regression (MLR)-based prediction, two different measures of chromatographic separation were used: (DP_i) and information content (I_i) with 5 and 10% error factors (E) [2].

Software implementations of the SOM algorithm are widely available. SOM_PAK 3.1 [13] is used for this work though it is not the only software implementation of SOM that is freely available. Calcu-

lations were made on a personal computer that runs under Linux. In all cases we have used rectangular quadratic SOM topology and bubble algorithm for network training. In order to test whether predefined conditions enable the network to settle down within a certain number of iterations we have done all network training in triplicate with different initial $\mathbf{w}_k$ component values. All training was made in two steps. The first step was SOM training with a wider neighborhood and faster learning-rate parameter and it was followed by the second training with decreased values of these parameters in order to avoid local minima problems. For the first training step, the number of iterations (n) was chosen to be 10^6 and the learning rate parameter was 0.05, while in the second step, the number of iterations was 10^7 and the learning rate parameter was 0.02. Neighborhood parameters ($h(i)$) chosen for SOM training are given in Table 1.

A Mathematica 4.0 multivariate statistics add-on program was used for PCA and MLR calculations. As a measure of prediction quality, the relative prediction error based on 'leave one out' method was used and implemented in Mathematica 4.0. In order to compare the prediction quality of the original MLR method used by the authors of Ref. [17] to SOM-based prediction quality, MLR calculations were performed and mobile phase compositions of chromatographic systems were used as input data while capacity factors were dependent variables. Based on these results, different chromatographic separation measures were calculated by KT1 software [2].

Table 1
Values of neighborhood parameter $h(i)$ chosen for SOM training

Map dimensions	1st training step	2nd training step
3×3	3	1
5×5	6	5
7×7	12	10
9×9[a]	15	12
11×11	20	18

[a] In order to increase the percentage of explained variance in Section 4.3, values of $h(i)$ for the 1st and 2nd training step were 17 and 14, respectively.

4. Results

4.1. Comparison of SOMs and dendrograms

The comparison of SOMs (feature maps are not shown) and dendrograms is presented in Fig. 2. It is based on topological distances between chromatographic systems in dendrograms (number of knots on the shortest path between two chromatographic systems) and SOMs (number of neurons on the shortest rectangular path between two chromatographic systems). SOM dimensions were varied between 3×3 and 9×9 neurons. To exclude noise effects on correlation, besides the Pearson correlation coefficient, the Spearman rank correlation coefficient has been calculated and the results are practically the same. This suggests that the observed correlations are not due to, or destroyed by noisy data. More importantly, there is no trend on the dependence of correlation coefficients connected with map dimensions. Though there is a 0.7 value of the correlation coefficient observed (Fig. 2(C)), correlations between dendrogram and SOM topologic distances are generally weak.

4.2. Applicability of SOMs in method development

Fig. 3 represents the dependence of the Pearson correlation coefficient between discriminating power (DP_i) and topological or Euclidean feature map distances of chromatographic systems from the best one among them (in terms of DP_i) on the number of neurons per map side (feature maps are not shown). These correlation coefficients give an insight in the applicability of SOMs for chromatographic method development. In cases where significant correlation exists it is justified to change the existing chromatographic system with a similar one in order to improve some chromatographic property, in our case resolution measured by DP_i. In this context, the term 'similar' stands for chromatographic objects whose topological or Euclidean distance in SOM is relatively small. The practical meaning of this finding is that one can replace old chromatographic conditions with a new set of conditions without losing or even for improving the value of some property because, for example the old system includes toxic or expen-

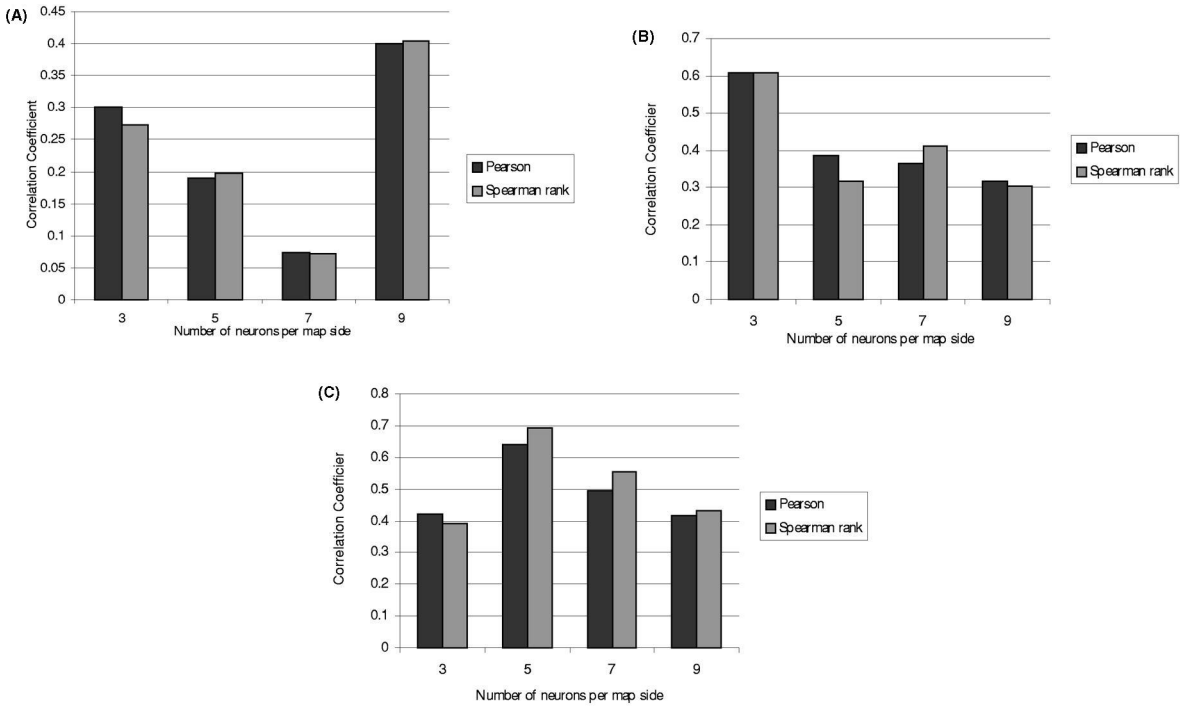

Fig. 2. Correspondence between dendrograms and SOMs described by Pearson and Spearman rank correlation coefficients. (A) Chromatographic objects taken for analysis are 11 TLC systems described by 11 R_F values of test compounds. Input data are taken from Tables 2 and 3 in Ref. [19]. (B) Chromatographic objects taken for analysis are nine HPLC systems described by 16 capacity factor values of test compounds. Input data are taken from Tables 2 and 3 in Ref. [18]. HPLC systems differ by flow-rate and/or mobile phase composition. (C) Chromatographic objects taken for analysis are 11 TLC systems described by eight R_F values of test compounds. Input data are taken from Tables 2 and 3 in Ref. [16].

sive reagents or some components of the system are not available. Those steps are important in early phases of method development. Chromatographers are very often in a position where they should make such changes but without evidence of correlation between similarity measure and certain chromatographic properties, mistakes are possible and the result is wasted time and high cost of method development. An advantage of the application of SOMs for this purpose is that when one has trained the SOM based on a certain number of experimental results and found the mentioned correlation, in order to find a similar set of conditions to the existing one, one need not to run one or more experiments. Rather one 'feeds' the existing map with the set of conditions one believes that are suitable and checks whether they are close to the existing set of conditions in SOM or not. This step saves a lot of effort because running this calculation is a lot easier than running a great number of experiments. Still SOM training requires a certain number of experiments.

As can be seen in Fig. 3, the correlation coefficient between similarity measure and DP_i varies significantly. In Fig. 3A, one cannot speak about significant correlations at all. This means that this map is not suitable for mentioned purpose. But in Fig. 3B–D, a correlation exists and these maps are suitable. There is also no trend in dependence of correlation coefficients connected with map dimensions. So, in Fig. 3A, better results probably can be obtained by changing map dimensions or by changing map topology. But care should be taken over the number of compounds and number of different sets of chromatographic conditions that are tested because

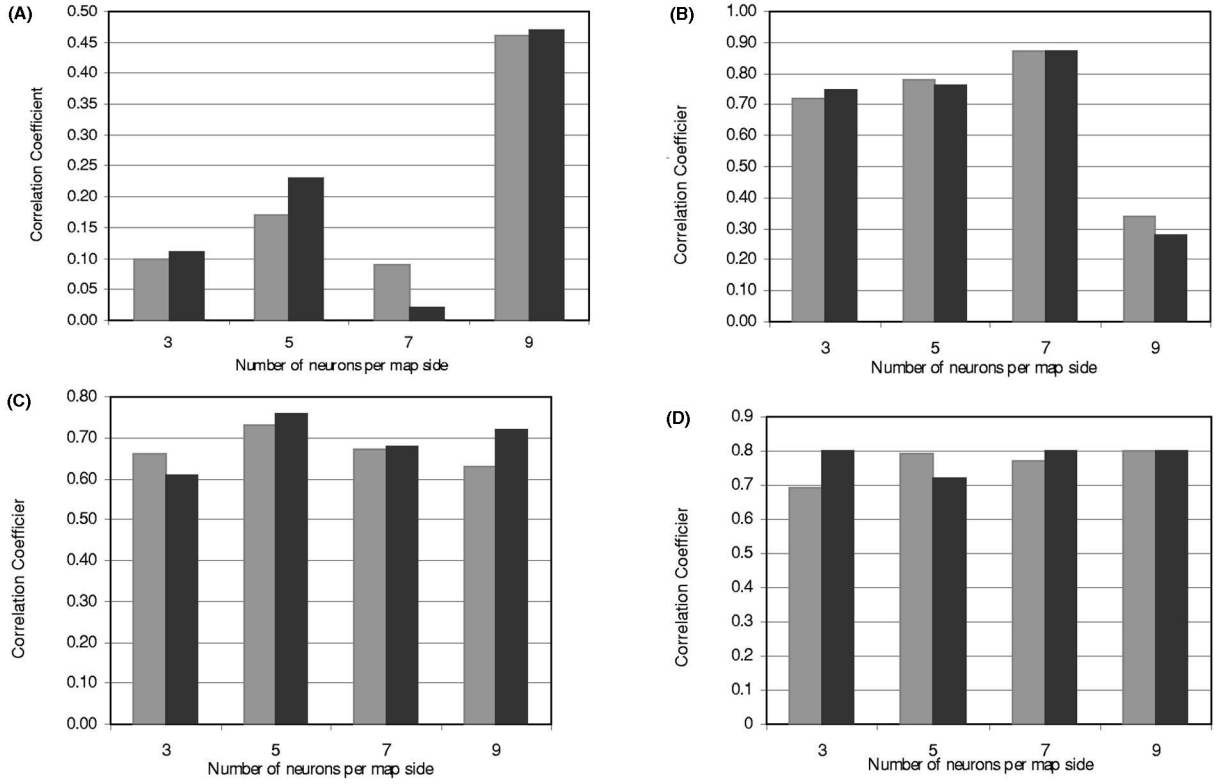

Fig. 3. Correlation between DP_i and trained SOM distances* of chromatographic objects. *Correlations of DP_i with Euclidean distances of chromatographic objects are presented by gray bars, while correlations with topological distances are presented by black bars. (A) Chromatographic objects taken for analysis are 11 TLC systems described by 11 R_F values of test compounds. Input data are taken from Tables 2 and 3 in Ref. [19]. (B) Chromatographic objects taken for analysis are 11 TLC systems described by eight R_F values of test compounds. Input data are taken from Tables 2 and 3 in Ref. [16]. (C) Chromatographic objects taken for analysis are 15 TLC systems described by 11 R_F values of test compounds. Input data are taken from Tables 2 and 3 in Ref. [2]. (D) Chromatographic objects taken for analysis are nine HPLC systems described by 16 capacity factor values of test compounds. Input data are taken from Tables 2 and 3 in Ref. [18]. HPLC systems differ by flow-rate and/or mobile phase composition.

map distances primarily depend on input data. Finally, depending on the property one wants to optimize one should decide what chromatographic system descriptors one would use as input. We have used capacity factors or retention times of test compounds since we were interested in resolution, but later in the text we shall show an example in which we use mobile phase composition as an input vector that describes a specific chromatographic system. The possibility to change independent or input variables is another good characteristic of the SOM approach because it gives some insight about variables that are critical for modeling of some separation quantity.

4.3. Comparison of SOMs and PCA

In order to compare the two mentioned classification tools, we have used data from Table 1 in Ref. [5]. Based on that source, we have trained an 11×11 SOM whose feature map is presented in Fig. 4. Fig. 1 from Ref. [5] corresponds to Fig. 4 in this paper. The reason for choosing this paper for comparison lies in the fact that authors of Ref. [5] have described only 86% of variance using the first two principal components in PCA presented in Fig. 1 [5]. In order to improve percent of variance explained by PCA, the authors should have included three or even four

60	58,62		17,79	5,18,27	19	46,59	26	80		2,55
	63		36		37	25		52		
21	6,77		70	57	1	61,69	49		3,45	
		4,64	75	74,78	66	16		29		84
22	23	15,30	31,38	13,35,67		20	68		48	7
73							28,81		34	
	72		9,14,33		24			56		8,53
							51			
65			71		83					47
		11		10		82		50		
32,39,41 ,42,43	40,44	54	12		85					76

Fig. 4. SOM-based classification of 85 RP-HPLC columns. Input data are taken from Table 1 in Ref. [5].

principal components but then graphical representation of the classification of HPLC columns they were dealing with would be impossible. It can be seen from Fig. 5A that four principal components are needed to explain 95% of the original data variance. So, those authors made their conclusions based on relatively low acceptance criterion. Fig. 5B shows the dependence of the percent of explained variance on map dimensions. It can be seen from Fig. 5B that the 11×11 map explains over 98% of variance. On the other hand, graphical representation is still easily reproduced in two dimensions (Fig. 4). So it can be concluded that, at least in this case, the SOM is better suited tool for the visualization and classification of chromatographic objects than PCA. Another reason for the taking mentioned paper for comparison is given in Fig. 2 of Ref. [5]. In order to obtain results which were more easily explainable the authors were in the position to exclude polar embedded C_{18} columns from analysis and even then they explained only 62% of the variance using the first two principal components. This step is not necessary and it is made in an ad hoc manner.

Without making this step in obtaining our results, they show a very nice correspondence to the results shown in Figs. 1 and 2 in Ref. [5]. Non-end capped

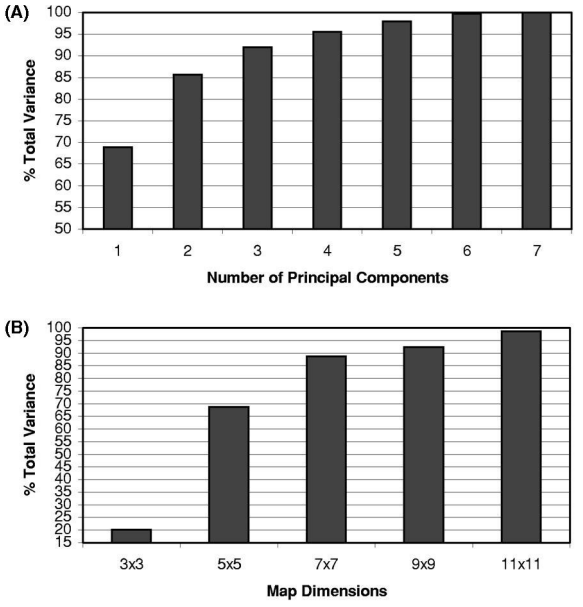

Fig. 5. (A) Dependence of explained variance on the number of principal components included in graphical presentation of PCA-based classification of 85 RP-HPLC columns. Input data are taken from Table 1 in Ref. [5]. (B) Dependence of explained variance on the map dimensions of the SOM-based classification of 85 RP-HPLC columns. Input data are taken from Table 1 in Ref. [5].

C_{18} columns, polar embedded C_{18} columns and C_{18} acidic columns of both types are classified in the same manner. Even this comparison shows one advantage of the SOM approach. In the original paper, all classifications were made by the authors' intervention. Namely, all partitions represented by circles in Figs. 1 and 2 in Ref. [5] were made without any theoretical explanation, while all partitions and subsets in SOM (Fig. 4) are in fact represented by positioning of members of subset in the same or in the neighboring neurons. In other words, all subsets are based on the difference between neuron vectors ($\mathbf{w}_k$) and chromatographic column vectors ($\mathbf{x}_i$). The only important difference in results is the classification of C_8 columns. While the authors of Ref. [5] classify them in more or less the same subset, our results are different. It could be misclassification in case, but from the experimental point of view, it is very often found that one can change one C_{18} column with a C_8 column and get more similar results then when one changes it with another C_{18} column depending upon the property one is analyzing. So it is still not clear what the explanation is for this result.

4.4. SOM-based prediction

Until now we have shown few applications of SOMs in the classification of chromatographic objects. But a very simple modification of this methodology leads to the application of SOMs for the prediction of, in our case, chromatographic quantities. Since the difference between ($\mathbf{w}_k$) and ($\mathbf{x}_i$) determines the location of the chromatographic object at the particular neuron, one can look at this difference as a probability measure of finding a particular chromatographic object in a particular neuron if this difference is normalized across all neurons for every chromatographic object. This normalization is the only step one needs to make to modify SOMs for prediction purposes. The products of the value of the quantity one is trying to predict with the corresponding normalized difference summed up over all chromatographic objects gives the prediction of that quantity for each neuron [12]. This approach is very practical. When one uses it for method optimization, one simply should find the neuron with the best-predicted value of the quantity one is trying to optimize and the vector of that neuron ($\mathbf{w}_k$) represents the set of the best chromatographic conditions one can get based upon this prediction. This is the advantage over the more classical back-propagated neural network approach where the prediction of the best conditions is not simple. The same statement is also true for the method robustness achievement.

Fig. 6 presents an example of SOM-based prediction. The data for SOM training were taken from Table 2 in Ref. [17] because the authors of this paper used a triangular mixture design for mobile phase optimization in order to achieve the best possible resolution of the test mixture and this opens the possibility to compare this approach with SOM. Mobile phase composition data were taken for SOM training. DP_i with 5% error factor was calculated based on capacity factor data from the same table. These data were used for SOM-based prediction described earlier. Finally, the results are presented as a map on to which the response surface has been overlaid (Fig. 6). Lighter areas represent higher

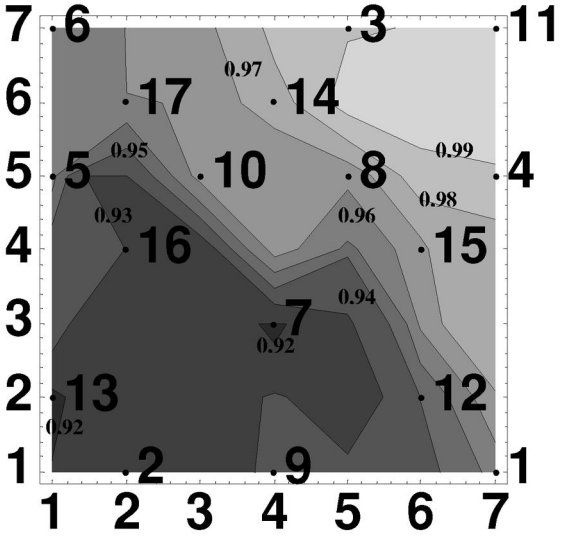

Fig. 6. SOM-based prediction of DP_i value. Input data, in this case mobile phase compositions, are taken from Table 2 in Ref. [17]. Lighter areas present higher values of DP_i. The DP_i surface has been overlaid on a 7×7 feature map (neuron coordinates are given on the left and lower side of the map; contour values are presented by small numerals placed near the corresponding contour itself; chromatographic mobile phase composition vector positions are presented by dots and large numerals).

values of DP_i calculated for data from Table 2 in Ref. [17]. A visual comparison of this figure with corresponding figures in the original paper (Fig. 2(C,D) shows a nice agreement. Still, it is a lot easier to visually analyze results in Fig. 5. The triangular mixture design is not easy to interpret even for ternary mobile phases and it is a particularly complex task when one deals with quaternary mobile phase and that is the case in the original paper [17].

Until now, we have used qualitative comparison to prove the applicability of SOM-based response surface analysis. To test the quality of prediction in a more quantitative manner we used the leave-one-out (LOO) method. Results of comparison between MLR-based prediction of chromatographic resolution measures used by the authors of Ref. [17] and SOM-based prediction are given in Fig. 7. It can be seen that MLR has a better prediction quality than SOM, and this is expected since SOM is not trained using any information on dependent variables, in this case chromatographic separation measures. Still, it can be stated that SOM-based prediction quality is

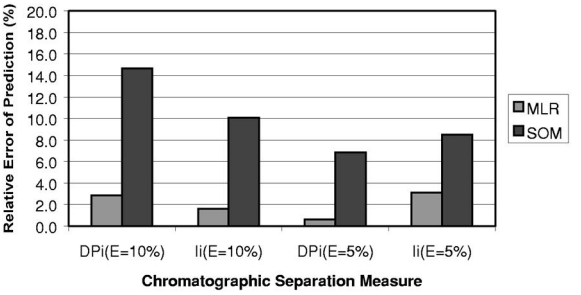

Fig. 7. Dependence of relative prediction error of the SOM-based and MLR-based prediction of different measures of chromatographic separation. Input data are taken from Table 2 in Ref. [17].

acceptable. Probably it can be improved by increasing the number of chromatographic test systems, adjustments of the number of iterations or by changing network topology.

5. Conclusions

The requirement for a quantitative approach to the classification of chromatographic objects for method development purposes has been presented. It is shown that all classifications do not need to correlate with the property one is trying to modify. Consequently, without evidence of correlation between 'similarity' measure and chromatographic property of interest there is no guarantee for any classification, whether it is SOM-based or not that it is well suited for a specific method development purpose.

SOMs can be applied in the classification of chromatographic objects, but also for response surface analysis and chromatographic property prediction. Comparisons show a better correspondence of SOM-based classification with PCA than with dendrograms. Quantitative measures and complexity of graphical representations show that SOMs are at least as well suited if not even better suited than the other two classification methods for method development purposes.

A reasonable value of relative error of prediction and an easy way to find the best predicted conditions are promising regarding use of SOM-based prediction in chromatography.

The SOM approach for method development is applicable to practically any type of chromatography.

We have shown its applicability in TLC and HPLC, but flexibility towards input and output quantities is so large that it allows adjustments to practically any type of chromatographic analysis.

References

[1] M. Mulholland, D.B. Hibbert, P.R. Haddad, P. Parslov, Chemometr. Intell. Lab. Syst. 30 (1995) 117.
[2] M. Medić-Šarić, Ž. Maleš, S. Šarić, Ž. Debeljak, J. Liq. Chromatogr. Relat. Technol. 22 (1) (1999) 83.
[3] P.M.J. Coenegracht, A.K. Smilde, H. Benak, C.H.P. Bruins, H.J. Metting, H. de Vries, D.A. Doornbos, J. Chromatogr. 550 (1-2) (1991) 397.
[4] C.F. Poole, S.K. Poole, Chromatography Today, Elsevier, Amsterdam, 1991.
[5] M.R. Euerby, P. Petersson, LC–GC, September (2000) 665–677.
[6] T.O. Kollie, C.F. Poole, J. Chromatogr. 550 (1-2) (1991) 213.
[7] M. Righezza, J.R. Chretien, J. Chromatogr. 556 (1-2) (1991) 169.
[8] J.V.S. Gobburu, W.L. Shelver, W.H. Shelver, J. Liq. Chromatogr. 18 (10) (1995) 1957.
[9] G. Sacchero, M.C. Bruzzoniti, C. Sarzanini, E. Mentasti, H.J. Metting, P.M.J. Coenegracht, J. Chromatogr. A 799 (1998) 35.
[10] H.J. Metting, P.M.J. Coenegracht, J. Chromatogr. A 728 (1996) 47.
[11] J. Gasteiger, J. Zupan, Angew. Chem., Int. Ed. Engl. 32 (1993) 503.
[12] B. Bienfait, J. Chem. Inf. Comput. Sci. 34 (4) (1994) 890.
[13] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, The Self-Organizing Map Program Package Version 3.1, Laboratory of Computer and Information Science, Rakentajanukio 2 C, SF-02150 Espoo, 1995.
[14] S. Haykin, Neural Networks, Prentice-Hall, Englewood Cliffs, NY, 1999.
[15] T. Kohonen, Self Organizing Maps, Springer, New York, 1995.
[16] M. Medić-Šarić, Ž. Maleš, Pharmazie 54 (1999) 362.
[17] P.M.J. Coenergracht, A.K. Smilde, H.J. Metting, D.A. Doornbos, J. Chromatogr. 485 (1989) 195.
[18] A. Rotar, F. Kozjek, M. Medić-Šarić, Acta Pharm. 43 (1993) 157.
[19] M. Medić-Šarić, G. Stanić, Ž. Maleš, S. Šarić, J. Chromatogr. A 776 (1997) 355.