Analusis, **1999**, *27*, 81-90
© EDP Sciences, Wiley-VCH

Clustering of infrared spectra with Kohonen networks

C. Cleva, C. Cachet and D. Cabrol-Bass*

GRECFO - LARTIC, Université de Nice Sophia-Antipolis, 06108 Nice Cedex, France

Abstract. The design of systems for spectral data interpretation requires clustering of chemical compounds based on their spectral characteristics. Kohonen networks have been shown to be efficient tools to achieve this clustering. These auto-organising systems perform a mapping between a high-dimensional variable space and a two-dimensional one. An application to infrared spectra of organic compounds is presented here. The non-supervised learning algorithm used allows classification of compounds by spectral characteristics without a priori knowledge. An analysis of the distribution of spectra on the resulting maps is used to build models for predicting the presence or absence of specific structural features. The performance of the models in recognising structural features is discussed and compared with the prediction of a multilayered feed forward network (MLFFN). Localisation of compounds wrongly classified by the MLFFN on the Kohonen maps allows to establish a link between the supervised and the unsupervised approaches.

Key words. Clustering – infrared spectra – neural network, Kohonen.

Introduction

Systems for spectral data interpretation need a formalisation of the relationships between the structure of compounds and their spectra. In spite of structure-spectra correlation studies using different methods, these relationships are poorly understood, especially in the case of infrared spectra. Such relationships are seldom simple or even linear, as reflected in the difficulty of building a base of rules for expert systems and the poor performance observed for a direct use of correlation tables [1]. The recent use of multilayered feed forward neural networks for the interpretation of spectral data shows good promise: the internal representations used by these networks are non-linear and are built up by a learning process based on examples. After the seminal works by Robb and Munk [2-3], many researchers explored the possibilities of these networks for the interpretation of infrared spectra [4-8]. However, the learning method used is a supervised learning process, and relies upon an existing structural classification. In fact, the networks learn to classify examples in a priori defined structural classes; in the case of infrared spectra, these classes are most often substructures corresponding to functional groups (alcohol, amine, ...). Moreover, the definition of structural classes based on the absence or presence of a specific structural feature may

*Correspondence and reprints.
Received February 20, 1998; revised October 30, 1998; accepted December 03, 1998.

Article available at http://analusis.edpsciences.org or http://dx.doi.org/10.1051/analusis:1999111

disregard possible interactions between the substructures, which could be important in infrared spectroscopy.

Other methods have been applied for clustering spectral data without the need for an existing structural classification. Most of these methods are linear (linear statistical models, k-nearest neighbours) and are poorly suited for infrared spectra classification [9]. Moreover, statistical methods have difficulties handling the high interdependence of spectral variables. Methods based on decision trees are also ill suited for spectral interpretation: they lead to sharp divisions of the population, and structural features are often spread over several clusters [10]. Kohonen networks seem not to suffer the problems encountered by these other methods. They are used in many domains for data classification [11-12]. These networks perform a projection of points from a high-dimensional data space onto a two-dimensional space. The learning process used is non-supervised: it requires no *a priori* information on classes and therefore classifies examples only by intrinsic characteristics (i.e. it performs a spectral classification of compounds). These networks have been little used for the classification of infrared spectra and have never been used directly as an interpretation system for general compounds. They have been used only for feasibility studies [13], selection of the training set for another interpretation system [14-16], or specific studies [17-18]. After a short introduction to Kohonen networks and its learning process, we present results on the clustering of infrared spectra of organic compounds from a commercial database [19]. The spectral classification obtained is used directly for the definition of models of some structural features, allowing a test of the use of this classification for spectral interpretation.

Method

Kohonen networks

We will recall only the basic principles of Kohonen networks in order to establish the vocabulary and notation used later. For a complete definition and detailed discussion about self-organisation, the reader might refer to reference [20]. The principal chemical applications of these networks are reviewed in [21].

Kohonen networks are composed of numerous elementary units which belong to either the input layer or the projection array. Units in the input layer are formal units, whose only function is to propagate examples (arranged as an input vector) to the units in the projection array. Units in the projection array are elementary processing units, regularly arranged in a two-dimensional array and connected to all input units. Therefore, each processing unit of the projection array receives the entire input vector. Associated with each processing unit is a weights vector of the same dimension as the input vector. When an example is presented to the network, that unit belonging to the projection array which is most similar to the example is selected and activated. It is usually referred as the winning unit (Fig. 1). The similarity

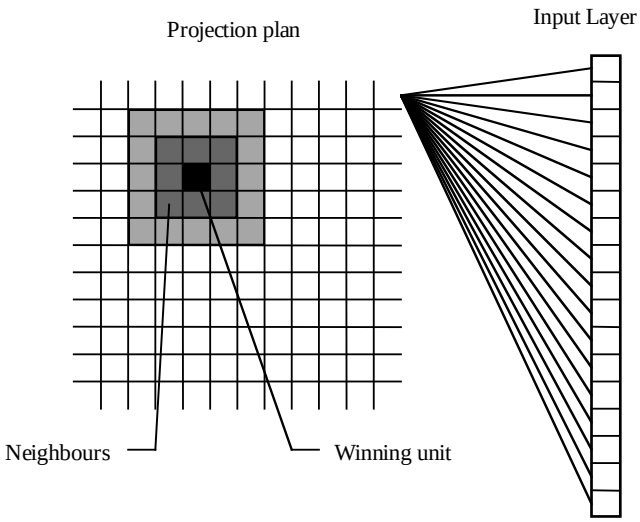

Fig. 1. Scheme of a Kohonen network. For the sake of clarity, only connections of one single unit have been drawn

of each unit to the example is computed as the inner product between its weights vector and the input vector of the example. The use of the complete spectra as input seems to perform better than other representations, even if, for spectral identification, peak matching methods show good results [22-23]. In the definition of a Kohonen network the inner product is used as a similarity measure. It has been shown that this measure is suitable for infrared spectra, assuming spectra are normalised [10].

The learning phase for Kohonen networks is based on a set of examples, called the training set. After initialisation of the weights vectors associated with the units of the projection array, the examples of the training set are presented one at time to the network and each activates a unique unit in the projection array. For each example, the weights vectors of the winning unit and its neighbours (the units surrounding the winning unit in the projection map) are modified in order to increase their similarity with the considered example. Let $x = \{x_j\}$ be the example. When this example is presented to the network, the weights w_{ij} of the unit i, neighbour of rank p (winning unit is the rank 0), are modified as:

$$\Delta w_{ij} = \alpha \cdot e^{-(p/N)^2} \cdot \left(x_j - w_{ij}\right)$$

$$\text{for } 0 \le p \le N \quad \text{with } 0 \le \alpha \le 1 \qquad (1)$$

$$\Delta w_{ij} = 0 \quad \text{for } p > N.$$

This process is repeated iteratively on the entire training set while parameters α and N, respectively the learning rate and the maximum rank of neighbourhood, are reduced after each example t by:

$$\alpha_{t+1} = \alpha_t.d\alpha \quad \text{and } N_{t+1} = N_t.dN \quad \text{with } 0 \le d\alpha, dN \le 1. \qquad (2)$$

Although the parameter *N* is defined as an integer (the number of ranks of units modified) a non integral value is used to allow a smooth decrease. The value is therefore rounded before its use in the computation.

The tuning of weights vectors of the winning unit and its neighbours is the basis of the self-organising process: during the learning process, the unit weights are modified so that similar examples would activate nearby units in the projection map, while dissimilar examples would activate distant units.

Once the network converges, or when values of parameters reach a threshold value, the learning process is stopped. After this training phase, the presentation of the whole training set of examples reveals clusters of units activated by similar examples. These clusters might be used for the definition of data classes (Fig. 2).

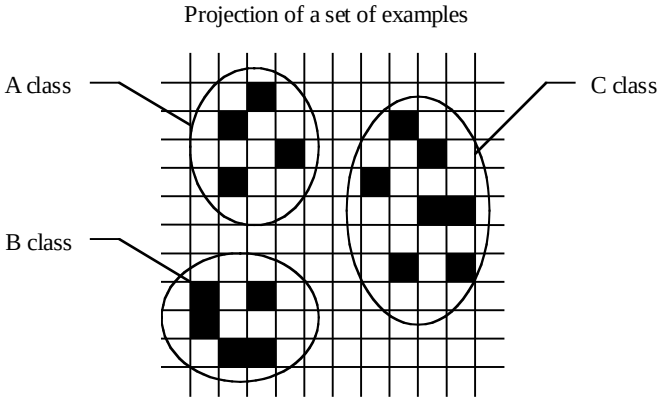

Fig. 2. Class definition by a Kohonen network.

Using a Kohonen network for prediction

Once trained, a network might be used in a prediction phase for the classification of examples which do not belong to the training set; this is called prediction. When the network processes an example, one unit of the projection array is activated, the location of which reflects the similarity between the considered example and the examples of the training set. If classes of examples have been found, the location of the winning unit directly indicates to which class or classes the example belongs. It should be mentioned that, in the prediction phase, examples not belonging to the training set are classified according to characteristics identified from the examples of the training set.

In fact, for each identified class a model based on the projections of the examples constituting the whole training set is generated. For each unit of the network, each model indicates whether the examples that activate that unit belong to a particular class. It is possible to define binary (yes - no), ternary models (yes - uncertain - no), or even fuzzy models (by assigning a membership value to each unit).

The performance of each model is measured by using a set of classified examples not belonging to the training set, called test set. Each example of the test set is presented to the network, and the model to be tested is applied. Once all examples of the test set have been classified by the model, the resulting classification is compared to the known classification, leading to a statistical measure of the performance of the model.

Application to infrared spectra

Infrared spectra are described by a *n*-dimensional vector consisting of the absorbance values at *n* different frequencies. The data space dimension *n* is typically high: for an infrared spectrum recorded between 800 cm^{-1} and 3600 cm^{-1} at a 1 cm^{-1} resolution, *n* equals 2801. This multidimensional representation of infrared spectra is difficult to analyse directly for classification, and the reduction of this dimension by statistical methods, such as principal component analysis (PCA), leads to a loss of information, and therefore to lower performance than Kohonen networks [18]. A Kohonen network, by projecting data of the *n*-dimension space on a two-dimensional space while keeping information on the relative distances between examples, might allow classification of the samples.

The projection map achieved by the trained network takes a useful form when the different classes are identified. Generally, the identification of clusters is equivalent to the identification of examples belonging to the cluster. Infrared spectra, however present a wide range of features, and the classes associated with these features overlap, making direct identification of clusters impossible. An exception occurs when the differences in spectra features are limited [17]. The classical approach is then to localise on the map the examples belonging to a predefined class. Classes may be defined by a physical property of the compound (for example, the identification of explosives vs. non-explosives [18]) or by the presence or absence of a particular structural feature in the compound [13-14,16]. For each class a map of the presence of the feature (or property) is so generated, and a model based on this map is produced. The localisation of a structural feature prior to the identification of classes leaves the learning process unsupervised: the map on which the structural feature is projected has been built only on the basis of spectral information, and the structural information is needed only afterwards to interpret the map.

Experiments

Learning parameters

All computations have been done on a PC computer. Neural networks have been simulated using SNNS [24] under Linux, while other processing has been accomplished using in-house programs.

Starting from a set of 10 000 spectra graciously put at our disposal by Biorad Sadtler [19], we selected 8 620

spectra of well-defined organic compounds by discarding salts, solvated, deutered, and organometallic compounds. From this resulting set, two subsets of 1 000 spectra each were generated by random selection, using distribution criteria (Tab. I). These independent sets are the same as those used for a previous study on a hierarchical system of feed-forward neural networks [8], allowing a comparison of the two approaches. The distribution criteria are used to check that the composition (proportion of compounds in each class) of the training and test sets is statistically representative of the composition of the whole database. Randomisation is assumed to ensure a good representation of each sub-class within the classes.

Infrared spectra, having a resolution of 12 cm^{-1} from 500 cm^{-1} to 3596 cm^{-1}, were coded as a vector of 259 absorbance values, and then normalised to a length of 1. Each weight vector of the units of the 30 × 30 array has been initialised by using a randomly assigned spectrum from the training set. The learning phase itself has been conducted with the following parameters values:

$$\alpha = 0.6 \quad N = 30$$

$$d\alpha = 0.999847 \quad dN = 0.999887.$$

These parameters were chosen by following the empirical rules used by different authors [13,20]. The size of the map (30 × 30 = 900 units) has been set considering the size of the training set (1000 examples) to ensure a wide distribution of the examples on the map with the presence of some empty units (i.e. with no spectrum projected on them). The initial value for the parameter N was set equal to the size of the map. The learning rate and number of iterations were chosen according to literature values. The learning computation has been reproduced during 30 epochs, one epoch corresponding to the presentation of the whole training set, or 30 000 iterations overall. Values for $d\alpha$ and dN have been chosen to ensure a smooth diminution of α and N from their starting value to, respectively, 0.006 and 1 during the learning process. An optimisation of those parameters would probably lead to a better classification and is under examination.

Projection of the whole set of spectra

After training is complete, the whole set of spectra constituting the training set is projected on the array of processing units. Out of the total number of 900 units (30 × 30), 329 are not activated at all, the remaining units being activated by a variable number of spectra ranging from 1 to 8. Figure 3 shows the number of spectra of the learning set which activate each unit of the projection array, leading to a so-called projection map or Kohonen map. This figure does not reveal distinct clustering of spectra which would allow the definition of well separated spectral classes. However, one can note marked effects on the border of the array, originating from the topological organisation of the array. Indeed, the units located inside the array have eight neighbouring units, while those located on the borders have only five and those on the corner only three. In order to avoid these distortions, one could use a torus instead of a two dimensional plane. Such a torus can be easily simulated by considering the units on each side of the array as neighbour of the units on the opposite side. On such a torus all units have the same number of neighbours, but because the maximum topological distance between two units is reduced by a factor of two, it would become necessary to use larger arrays of units. This increase in the arrays size would lead to a much longer training time, a solution which was rejected.

Table I. Number of compounds in each structural class.

Compounds	*Quantity in the training set*	*Quantity in the test set*
hydroxylic	359	359
carboxylic	490	490
amino	344	359
benzenic	441	509
ethylenic	209	209

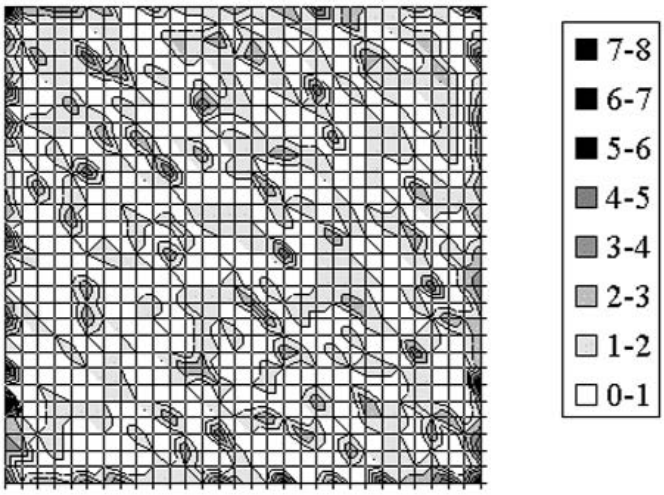

Fig. 3. Map of the number of spectra of the learning set projected on each unit.

Localisation of spectra of compounds having a particular substructure

The same type of maps can be obtained by counting, for each unit on the projection map, the number of spectra of compounds having a particular substructure. The resulting maps are characteristic of the clustering of spectra of this particular structural class. Such maps have been generated for the five general classes of compounds: hydroxylic, carboxylic, amino, benzenic and ethylenic compounds. As an example, figure 4 shows the maps of the projection of spectra of hydroxylic compounds.

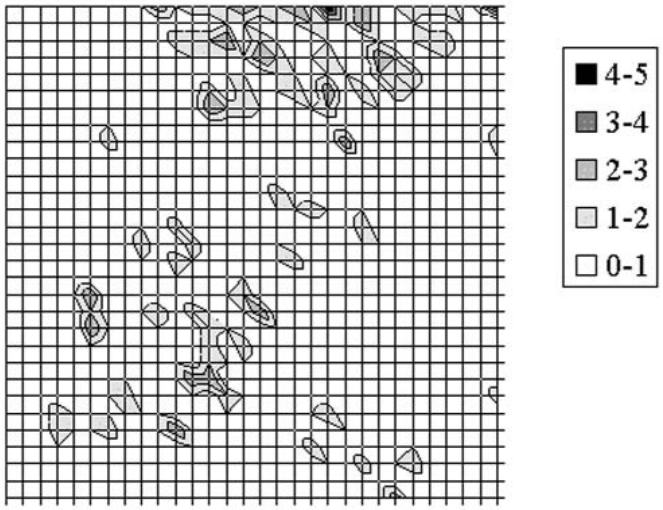

Fig. 4. Map of the number of spectra of hydroxylic compounds projected on each unit.

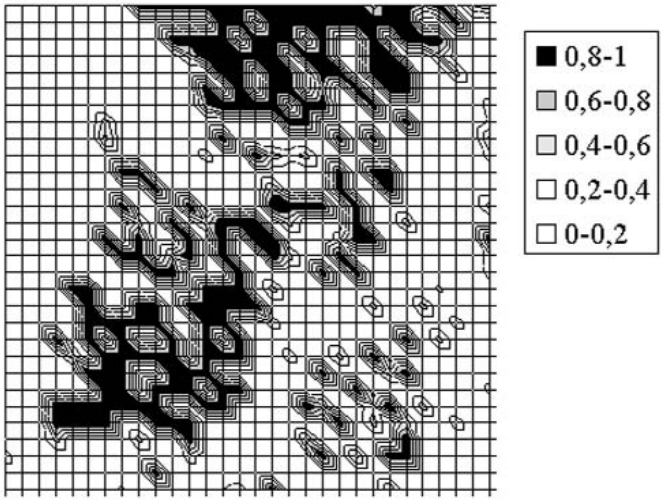

Fig. 5. Map of the ratio of spectra of hydroxylic compounds projected on each unit.

Defining the boundaries of zones is difficult on maps constructed using the number of spectra on each unit of the projection plane. It is preferable to represent for each unit the ratio of the number of compounds having the particular substructure to the total number of compounds whose spectra are projected on this unit. Figure 5 shows the resulting map for hydroxylic compounds.

On this type of map, one can identify one or more characteristic zones for each structural class. Units containing a value of zero within the limits of a characteristic zone correspond in the majority of cases to units on which no spectra are projected at all. In other words, rarely are spectra of compounds which are not members of the considered structural class projected separately on a unit located inside the characteristic zone of this class. Moreover, a high fraction of compounds having the considered substructure have spectra which are projected in "pure zones" made of units for which all the projected spectra correspond to compounds of the same structural class (see Tab. II). The measurement of this ratio, combined with the visual inspection of the obtained zones, allow the assessment of the quality of the clustering of spectra for compounds in the same structural class. The diffusiveness of the zones of clustered spectra on the Kohonen map relates to the spectral features specific to the considered structural class. If these features are unimportant compared to others, then the spectra will be poorly grouped, as it is the case for ethylenic compounds (Fig. 6). One can note the number of units activated by compounds belonging to a class depends upon the percent of those compounds in the training set: with about 36% of hydroxylic compounds in the training set those compounds are projected on about 29% of the Kohonen map, while the carboxylic compounds (49%) occupy about 36% of the Kohonen map.

Table II. Clustering of compounds within each class.

Compounds	*Percent of compounds in "pure zones"*
hydroxylic	77.4
carboxylic	86.5
amino	70.9
benzenic	70.1
ethylenic	48.3

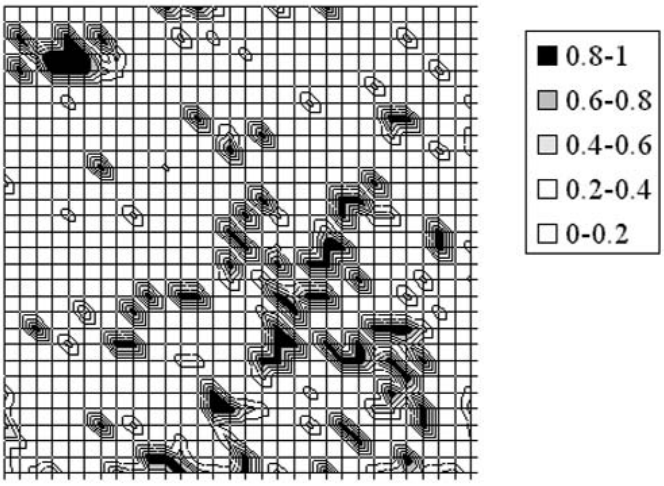

Fig. 6. Map of the ratio of spectra of ethylenic compounds projected on each unit.

We have constructed another type of map by computing for each unit i the sum of the inner products of its weights vector w_i with the spectral vectors x of all compounds

having a particular substructure (i.e. belonging to class A). Dividing this sum by the number of compounds in this structural class, one obtain $\overline{S}_i^A$, the mean similarity of the considered unit with the collection of spectra of the class A.

$$\overline{S}_i^A = \frac{1}{\mathrm{card}(A)} \sum_{x \in A} \sum_j w_{ij} \cdot x_j. \quad (3)$$

Through this process a similarity map is obtained for each substructure. On the map, units having high values are the most representative of the considered class. Figure 7 shows the map obtained for hydroxylic compounds. The iso-similarity curves of each map look like those on the map obtained by projection of the units activated by compounds having the considered substructure. The observed differences originate from the variability of spectra within the class itself: the higher the variability of the spectra in a class, the more the summation operation will level the similarity map.

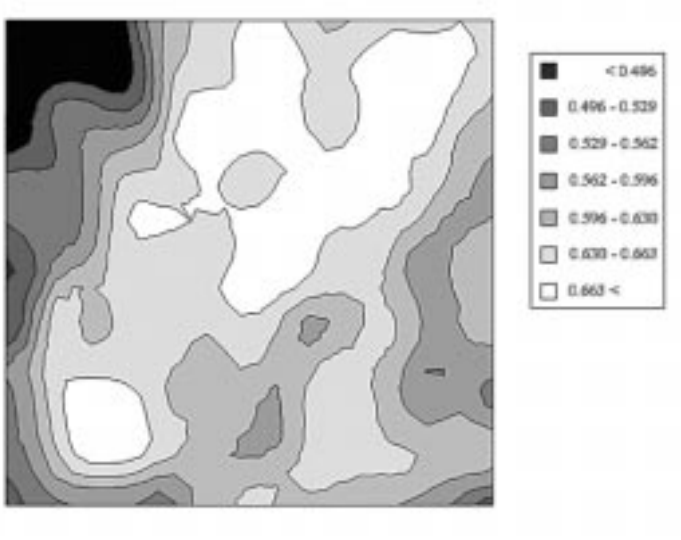

Fig. 7. Sum distance map of spectra of hydroxylic compounds of the learning set. The higher the inner product, the closer from spectra of hydroxylic compounds is the considered unit.

Model construction and performance evaluation

Model construction

A model is defined as an area of the unit array which is associated with a particular substructure class. For each previously defined substructure class, we have constructed two models. The first one is based on the winning units maps, while the second is derived from the similarity maps.

Models based on the winning units are built using the maps of the substructure ratio resulting from projection of the spectra of compounds belonging to learning set. First, the initial cluster(s) for a substructure class is made of units which are activated solely by spectra of compounds which contain the particular substructure (pure zones). Next, units which lie inside these zones (that is to say surrounded by units of the initial cluster) but which are not activated by any spectra are incorporated in the model. Those units are not activated by any spectra of the training set but might be activated by spectra of compounds of the same class belonging to the test set. Figure 8 shows the weights vector of two neighbour units, one activated by the spectra of hydroxylic compounds, and the other unactivated by any spectra. After convergence, the weights vector associated with each unit lies close to the spectrum which activated this unit. Therefore, it is not surprising that the weights vectors of neighbouring units are similar. In practice, units which are surrounded by at least 4 activated units are incorporated into the model (this number has been chosen empirically). Figure 9 shows the model for hydroxylic compounds resulting from the map of figure 5.

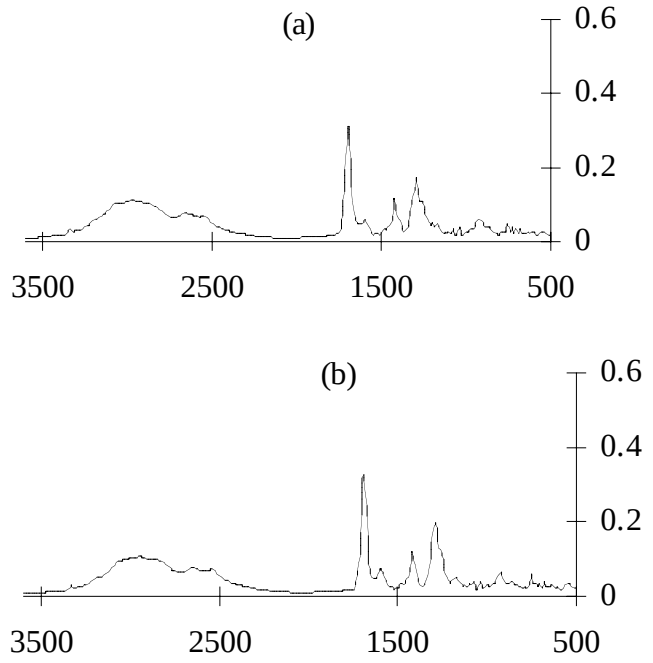

Fig. 8. Comparison of weights vectors from neighbour units. The spectrum of the m-chloro benzoic acid activates the node (a), while no spectra of the training set activates the node (b).

The second model is obtained from the similarity map of each substructure class. For each unit *i*, the mean similarity to the class A ($\overline{S}_i^A$) is used to decide whether the unit is to be incorporated into the model of this class. For each unit three states exist to define its association with a class: positive, negative or undefined. For this purpose it is necessary to fix two threshold values per class: the Accept Level (AL) and the Reject Level (RL). For a particular unknown compound, its spectrum activates one unit of the projection array. If the $\overline{S}_i^A$ computed for this unit is higher than AL one considers the compound to contain the considered substructure (i.e. belong to the class A). If $\overline{S}_i^A$ is lower than RL,

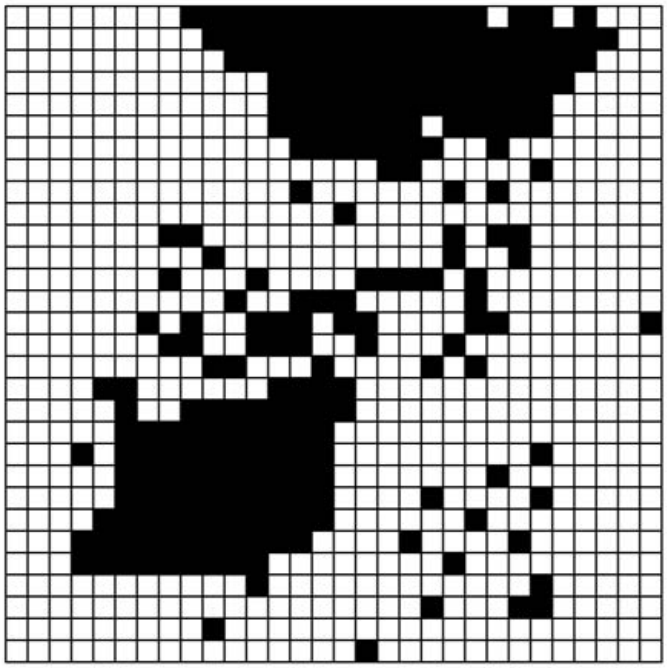

Fig. 9. Model for the class of hydroxylic compounds derived from the winning units map. Units in black correspond to spectra of compounds for which the hydroxyl substructure is considered as present.

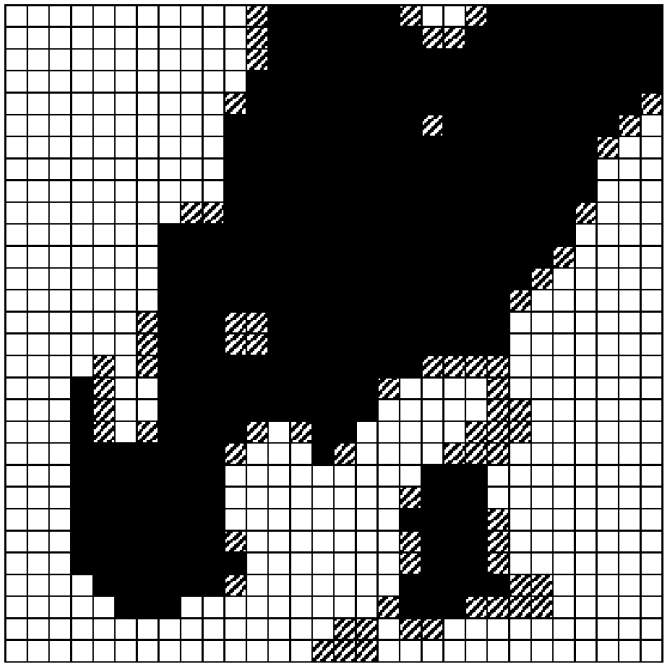

Fig. 10. Model for the class of hydroxylic compounds derived from the sum distances map. Units in black correspond to spectra of compounds for which the hydroxyl substructure is considered as present. Striped units correspond to undecided cases.

then the opposite conclusion is drawn. If $\bar{S}_i^A$ is between RL and AL it is not possible to decide whether the considered substructure is present. The threshold values AL and RL are obtained by an optimisation procedure using the simplex method in order to maximise the overall performance of the model on the whole training set. Figure 10 shows the model obtained by this method for hydroxylic compounds.

Performance evaluation

The various resulting models have been used to classify the compounds in the test set. For each compound in the set, if the unit activated by its spectra belongs to one or several models, then the compound is classified as a member of the corresponding structural class or classes. The responses obtained in this way by each model are compared to the effective (true) class membership of the compound. Overall network performance with regard to classification is evaluated by computing the number falling into each of four categories for the whole test set: compounds having the considered substructure (Present) and correctly classified (PcP: Present classified as Present), or wrongly classified (PcA: Present classified as Absent); compounds which do not have the considered substructure (Absent) and correctly classified (AcA: Absent classified Absent) or wrongly classified (AcP: Absent classified Present). Models based on similarity maps also produce a number of unclassified compounds, but these numbers do not directly affect the calculation of performance indices. The following indices are used to evaluate and compare network performance. For a complete discussion of the reliability of these indices see reference [25]. Each index is relative to a specific substructure class.

$$Pf = \frac{PcP}{Np} \tag{4}$$

$$Af = \frac{AcA}{Na} \tag{5}$$

$$Qpr = \frac{PcP}{PcP + AcP} \tag{6}$$

$$Qar = \frac{AcA}{AcA + PcA} \tag{7}$$

$$GQ = \frac{PcP + AcA}{Np + Na} \tag{8}$$

$$EQr = \frac{GQ - St}{1 - St} \tag{9}$$

with $St = 1 - 2Pi + 2Pi^2$

and $Pi = \frac{Np}{Np + Na}$. (10)

Indices *Pf* (fraction of Present found) and *Af* (fraction of Absent found) represent respectively the fraction of those compounds containing (Present) and not containing (Absent) the considered substructure that where correctly classified. *Np* and *Na* are the number of those compounds.

Indices *Qpr* and *Qar* represent respectively the quality of the "Present" and "Absent" responses, that is to say the percentage of correct answers when the response is given as "Present" and "Absent".

The global quality *GQ* is the ratio of correct responses to the total number of examples. *EQr*, named "Extra statistical Quality of responses" is a measure of the relative improvement of the global quality over that which would have been expected by chance given the statistical distribution of the population. Thus *EQr* is a measure of the fraction of improvement that has been achieved by the model over random selection. The higher the value of *EQr*, the greater the discrimination achieved by the network. Note that one could use $St = \max(Pi, 1 - Pi)$ instead of equation (9) to calculate the *a priori* probability of obtaining a correct response, resulting in slightly different values for *EQr*. Although the global quality *GQ* is very frequently used in the literature, *EQr* is to be preferred to estimate the model quality because it alleviates from the bias introduced by unbalanced distribution.

Results comparison

The results obtained for each model are reported in table III.

For all the structural classes studied, the model derived from the similarity map gives a lower *EQr* than the model based on the winning units. The lower performance can be explained by the variability of the spectra within each structural class. Indeed when a structural class includes several subclasses, and if the spectra of the compounds of these subclasses differ significantly, the summation of similarity made to build the similarity maps has an important levelling effect. In such cases the models based on the similarity maps are less discriminating than those constructed from the winning units. Comparisons of *EQr* values between the models must be made with caution because other factors could have an effect on the values. Nevertheless, the higher differences observed in the case of hydroxyl and carbonyl classes without doubt result from the high variability of spectra of compounds belonging to these classes. Figure 11 shows the distribution of projection of spectra of hydroxylic compounds on the unit array. This projection reveals the existence of separate subclasses which explains the differences between the models for this structural class (see Figs. 9 and 10).

The variability of spectra within the hydroxylic class is exemplified by figure 12. Units near the centre of clusters (e and f in Fig. 11) are activated by characteristic spectra of the corresponding sub-class (e: benzyl alcohol *m*-methyl and f: *m*-chloro benzoic acid, not represented). Units on the border of the main clusters (units a, b, c and d in Fig. 11) are activated by spectra presenting both functional characteristics (a: cycloheptanecarboxylic acid, 1-hydroxy and d:

Table III. Performance of the classification models for each structural class. Performances of the feed-forward layered network are given for the sake of comparison.

Models based on presence maps

Compounds	**Pf**	**Qpr**	**Af**	**Qar**	**EQr**
hydroxylic	63.5	84.4	93.5	82.1	62.4
carboxylic	78.6	92.3	93.7	82.0	72.6
amino	52.1	74.8	90.2	77.1	48.9
benzenic	56.4	81.8	87.0	65.8	42.8
ethylenic	13.4	38.4	94.3	80.5	31.7

Models based on similarity maps

Compounds	**Pf**	**Qpr**	**Af**	**Qar**	**EQr**
hydroxylic	67.7	63.3	73.3	84.8	37.6
carboxylic	72.2	79.5	71.0	78.9	43.2
amino	51.0	76.9	84.7	82.2	40.5
benzenic	83.3	68.2	54.8	82.0	38.6
ethylenic	7.18	25.0	91.8	79.5	21.7

Hierarchical neural network [7]

Compounds	**Pf**	**Qpr**	**Af**	**Qar**	**EQr**
hydroxylic	83.3	87.4	93.3	91.0	77.6
carboxylic	91.8	96.6	94.3	94.7	86.2
amino	82.7	87.6	92.0	91.3	75.4
benzenic	86.4	84.6	79.0	92.8	65.6
ethylenic	42.6	64.0	93.6	86.1	48.3

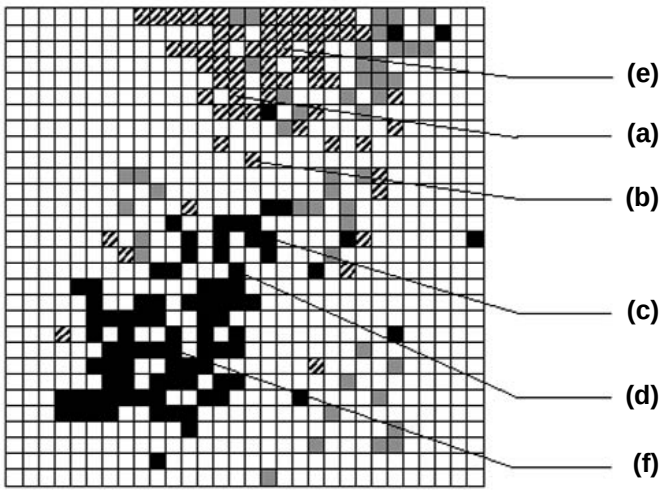

Fig. 11. Analysis of projections of spectra of hydroxylic compounds. Units in black correspond to carboxylic acids, striped units to aliphatic alcohols. Other hydroxylic compounds (in grey) correspond principally to phenols.

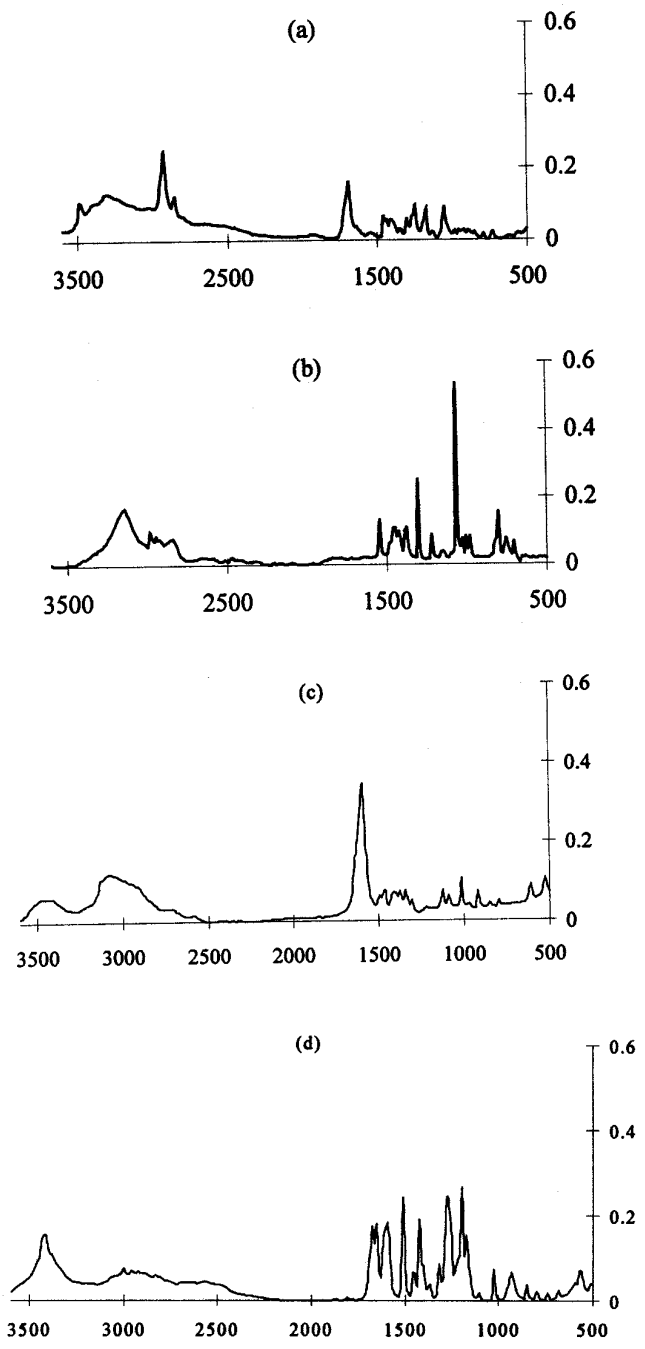

Fig. 12. Spectra of (a) cycloheptanecarboxylic acid, 1-hydroxy; (b) pyrazole-1-methanol, 3-5-dimethyl; (c) hydracrylic acid, 2-amino, L-/plus/-; (d) cinnamic acid, 4-hydroxy, 3-methoxy. See figure 11 for localisation.

cinnamic acid, 4-hydroxy, 3-methoxy) or influenced by other functional groups (b: pyrazole 1-methanol, 3,5-dimethyl and c: hydracrylic acid, 2-amino).

For the other three structural classes, the difference in *EQr* is smaller. An analysis of the values of the other

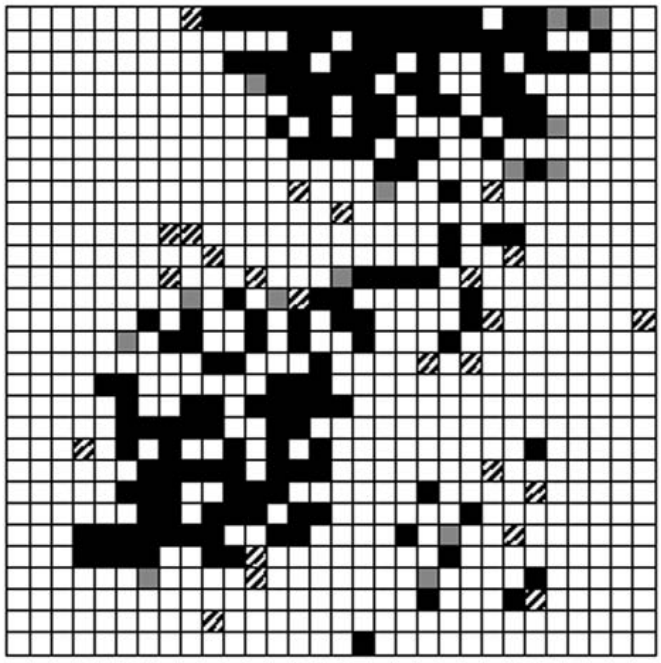

Fig. 13. Analysis of the responses for hydroxylic compound obtained by a layered network projected on the Kohonen map. Correct positive classifications are represented in black, false positive in grey, false negative in black and white striped.

performance indices and of the models themselves reveals other differences. The two models obtained for ethylenic compounds differ noticeably and their performance values are low. This poor performance originates from an ill-defined clustering of spectra reflecting intra-class variability but lacking the visible formation of sub-classes (Fig. 6). The two models for the two other classes are similar. In the case of benzenic compounds, one of the model is too restrictive (leading to many false negative responses) while the other is too inconclusive (leading to many false positive responses) but the resulting values of *EQr* are comparable. The small difference is an indication that the spectra of compounds belonging to the class and those which are not members of the class do not differ significantly. Models for the amino compounds are also similar while the similarity map shows a levelling effect. However this levelling effect is less important than in the cases of hydroxyl and carbonyl compounds, the result of a lower variability of the spectra.

Comparison of model performance with those achieved by a layered network

The performance of the models based on winning units has been compared with those achieved by a hierarchical layered network which has been trained to recognise the same structural classes on the same learning set [8] (see Tab. III). The results obtained with the Kohonen network are poorer than those obtained with the layered network. The values of extra statistic quality *EQr* for all structural classes obtained by the

two types of networks have a correlation coefficient of 0.94 leading to the conclusion that, despite their different architecture and classification method, they both rely on similar basis. This finding, already stated by Melsen et al. [13], is reinforced by the localisation on the Kohonen map of the projections of the spectra of compounds from the learning set which have not been properly classified by the layered network. Figure 13 shows these projections for spectra of misclassified hydroxylic compounds. Errors of the layered network occur for spectra which are projected on the borders of the Kohonen model or in areas in which the clustering of examples is indistinct.

These remarks lead to the assurance that an MLFFN would make an excellent classification of spectra in classes previously defined by the means of a Kohonen network. A confirmation experiment is under investigation.

Conclusion

The classification of infrared spectra by Kohonen networks provides useful insight for the analysis of substructure-spectra correlation and for construction of systems to interpret spectra, in particular in view of structure elucidation. For a large collection of compounds of different structure, the resulting spectra classification is fuzzy and can not be used directly. However the analysis of the spectra projections on the Kohonen map of compounds containing a particular substructure allows the definition of models for this structural class. Models based on the location of winning units lead to a better classification performance than those derived from the sum similarity. The difference in performance is more noticeable when the spectra of a general class show a higher variability, in particular when they can be distributed between several sub-classes having distinct spectral characteristics.

Even though the classification performance of Kohonen networks could be improved, notably by optimising the learning parameters or possibly by using a larger learning set, it seems clear from our results that for this specific application, they are outperformed by feed forward layered networks. The difference lies in the fact that Kohonen networks use a global similarity calculated on the whole spectrum, while layered networks can use various combinations of local spectral characteristics (a kind of “XOR” classification), allowing a release from spectral variability in spectral regions which are not specific to a structural class.

Despite these limitations, spectral classification by Kohonen networks are useful for the definition of substructural classes to be identified by feed forward layered networks. As a mater of fact, analysis of the locations on the Kohonen network of spectra erroneously classified by the layered networks indicate the close relationships between the two. Therefore, defining spectral classes and sub-classes using the Kohonen networks should lead to the construction of layered networks having improved performance.

Acknowledgement

Dr. Richard Cornelius (Lebanon Valley College, Annville PA, USA) is gratefully acknowledged for his help with the linguistic revision of the manuscript.

References

1. Affolter, C.; Baumann, K.; Cler,c J. T.; Schriber, H. *Mikrochim. Acta. [Suppl.]* **1997**, *14*, 143-147.
2. Robb, E. W.; Munk, M. E. *Mikrochim. Acta (Wien)* **1990**, *1*, 131-155.
3. Munk, M. E.; Madison, M. S.; Robb, E. W. *Mikrochim. Acta (Wien)* **1991**, *2*, 505-514.
4. Meyer, M.; Weigelt, T. *Anal. Chim. Acta*, **1992**, *265*, 183-190.
5. Ricard, D.; Cachet, C.; Cabrol-Bass, D.; Forrest, T. P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 202-210.
6. Smits, J. R. M.; Schoenmakers, P.; Stehmann, A.; Sijstermans, F.; Kateman, G. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 27-39.
7. Klawun, C.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 69-81.
8. Cleva, C.; Cachet, C.; Cabrol-Bass, D.; Forrest, T. P. *Anal. Chim. Acta* **1997**, *348*, 255-265.
9. Burns, J. A.; Whitesides, G. M. *Chem. Rev.* **1993**, *93*, 2583-2601.
10. Frankel, D. S. *Anal. Chem.* **1984**, *56*, 1011-1014.
11. Zupan, J.; Novic, M.; Li, X. Z.; Gasteiger, J. *Anal. Chim. Acta* **1994**, *292*, 219-234.
12. Wienke, D.; Hopke, P. K. *Anal. Chim. Acta* **1994**, *219*, 1-18.
13. Melsen, W. J.; Smits, J. R. M.; Rolf, G. H.; Kateman, G. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 195-204.
14. Novic, M.; Zupan, J. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454-466.
15. Wu, W.; Walczak, B.; Massart, D. L.; Heuerding, S.; Erni, F.; Last, I. R.; Prebble, K. A. *Chemom. Intell. Lab. Syst.* **1996**, 35-46.
16. Zupan, J.; Novic, M.; Ruisanchez, I. *Chemom. Intell. Lab. Syst.* **1997**, 1-23.
17. Caceres-Alonso, P.; Rodriguez-Galan, R.; Garcia-Tejedor, A. Proc. 7th Int. Conf. Near Infrared Spectrsc. 1995; pp 393-398.
18. Daniel, N. W.; Lewis, I. R.; Griffiths, P. R. *Mikrochim. Acta. [Suppl.]* **1997**, *14*, 281-282.
19. Copyright 1980-1993 Bio-Rad Laboratories Inc., Sadtler Division, 3316 Spring Garden Street, Philadelphia, PA 19104, USA. All right reserved.
20. Kohonen, T. Self-Organization and Associative Memory, 3rd edition, Springer, New-York, 1984.
21. Gasteiger, J.; Zupan, J. Neural Networks for Chemists, VCH, Weinheim, 1993.
22. De Ruig, W. G.; Weseman, J. M. *J. Chemom.* **1990**, *4*, 61-77.
23. Visser, T.; Luinge, H. K.; Vandermas, J. *Mikrochim. Acta. [Suppl.]* **1997**, *14*, 287-288.
24. Stuttgart Neural Network Simulator, Institute for Parallel and Distributed High Performance Systems, University of Stuttgart. Disponible à ftp.informatik.uni-stuttgart.de.
25. Sbirrazzuoli, N.; Cachet, C.; Cabrol-Bass, D.; Forrest, T. P. *Neural Comput. Applic.* **1993**, *1*, 229-239.