Application of Self-Organizing Map Algorithm combined with Structuring Index to characterize strawberry variety aroma by SPME/GC/MS

J.L. GIRAUDEL, V. DE BOISHEBERT, M. MONTURY.
EPCA - Laboratoire de Physico et Toxico Chimie des Systèmes Naturels
Université Bordeaux 1 - CNRS (UMR 5472)
BP 1043, 24001 Périgueux Cedex
FRANCE

Abstract: -The Kohonen Self-Organizing Map (SOM) is one of the most well-known neural network with unsupervised learning rules; it performs a topology-preserving projection of the data space onto a regular two-dimensional space. So, SOM is considered as a powerful tool for data mining and SOM can be recommended for studying large, high-dimensional data sets.
This method has already been successfully applied to discriminate strawberry variety aroma basing on the use of their chemical fingerprints delivered by SPME/GC/MS analysis. Once the learning step achieved, a lot of methods can be used to visualize the SOM. For example, showing the values of one chemical compound in each map unit, the component plane visualization allows to show the part of each variable in the total organization of the map but also in some of its areas. However, with high-dimensional data, a lot of maps have to be considered and the task can become fastidious and even quite impossible to achieve. The purpose of this paper is to propose a computational method to determine the most relevant variables for structuring the obtained map.
So, the concept of Structuring Index (SI) has been used in order to determine the most relevant chemical compounds for structuring the SOM obtained in the case of these strawberry aroma analyses. In a future extension, SI could provide an efficient tool for pre-processing data, for instance in order to reduce the number of variables used in the input layer of a multilayer artificial neural network.

Key-Words: -self-organizing map; neural networks; structuring index; aroma SPME analysis; strawberry variety aroma.

1 Introduction

In order to provide an efficient and running analytical tool to strawberry plant breeders who have to characterize and compare the aromatic properties of new cultivars to those already known, a chemical method using head-space solid phase micro-extraction (HS-SPME), gas chromatography - mass spectrometry analysis (GC-MS) and statistical treatment of the results by Self Organizing Map (SOM), has been recently developed by the authors [1]. According to this approach, this method based on the relative determination of 23 of their chemical constituents allowed to characterise aromas of 17 strawberry varieties and to fully discriminate them. In this frame, the efficacy of such an unsupervised learning system applied to the matrix of results, proved very promising. Actually, a 24-unit map revealed sufficient for obtaining the full discrimination of the 70 analysed samples, gathering all samples from the same variety in the very same unit and no unit containing several varieties. Moreover, analysis of unknown samples belonging to the panel of the studied varieties, were correctly classified by the whole method and projected in the unit corresponding to the right variety, so providing a recognition function to this analytical approach.
Recently, Dutta et al. [2] used the same approach in order to discriminate different qualities of tea by using a similar process. To our knowledge and except this very recent example, no natural noses through any sensorial analysis methods [3] or artificial noses through any physical or chemical analysis systems [4] are presently able to realise such a discriminative classification and recognition performance. In these conditions, the described method dealing with the aroma chemical constitution appears of real interest for plant breeders for whom the discriminative classification of existing varieties is the inescapable first step for positioning a new cultivar comparatively to parent varieties and all previously existing others.
Another aspect of this data mining may concern the way of visualizing the maps issued from the SOM algorithm [5]. For example, the role played by each chemical constituent as a variable parameter of the data set for structuring the computed map, can be pointed out by an adapted treatment. In case of large data matrices and if the objective is to determine which constituents are the most relevant for the map discrimination effect,

corresponding running series may turn fastidious and often unachievable. Actually, the 23 chemical compounds used by Urruty et al. [1] were selected according to the literature as being the main compounds in terms of quantity and sensorial effect. In fact, more 300 constituents have already been identified in strawberry aroma by chemical analysis [6]. The objective of the present study is to exemplify a computational method for determining the most relevant constituents that have to be analysed as structuring agents of the map, in order to afford the same efficacy to the method. So, 100 chemical constituents were used and a structuring index (SI) has been defined for all analysed chemical compounds to study a potential reduction of the number of variables.

2 Materials and Method

2.1 Chemical Analysis: Main Features

To implement the models, the dataset came from the chemical analysis of 8 strawberry varieties (Table 1) provided by the producer in charge of the selection program (CIREF, 24120 Prigonrieux, France).

Table 1
Presentation of the Selected Varieties

Code	Variety
A	CF129
B	Ciflorette
C	Cigaline
D	Cilady
E	Ciloé
F	Cireine
G	Cigoulette
H	Capitola

The analytical method is divided in two steps. The first one consisted to extract the volatiles from the strawberry juice by SPME (Solid Phase Micro Extraction). SPME, developed by Pawliszyn at the beginning of the nineties [7], is a multi-analyte extraction technique that requires no solvents and provides results over a wide range of analyte concentrations. The procedure is to introduce a fused silica fiber into the headspace above the sample in order that the analytes were adsorbed by the fiber coating. Compared to all the commercially available fibers, the 65 m polydimethylsiloxane / divinylbenzene fiber (Supelco) revealed a good efficiency to extract the strawberry volatile compounds. The second step consisted to separate and identify the extracted volatiles. For this operation, a gaz Chromatograph Trace GC (ThermoQuest, Les Ulis, France) coupled with a Polaris Mass Spectrometry (ThermoQuest, Les Ulis, France) was used. After 30 minutes of extraction, the fiber was transferred into the chromatograph injector where the analytes were thermally desorbed at 250°C during 10 min. Then the analytes were separated through a capillary column MDN-5S 30m x 0.25mm x 0.25 m and detected by a mass spectrometer ion trap. Results obtained by GC/MS were chromatograms as this presented in Fig. 1 that was characteristic of the corresponding sample and considered as its aromatic chemical signature.

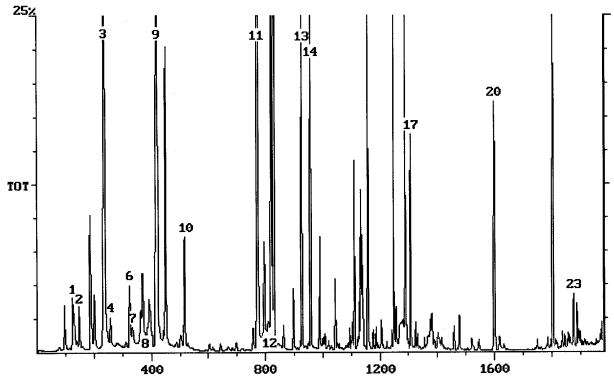

Fig. 1: chromatogram from Capron Royal

For this study, in addition to the 23 chemical compounds previously used by Urruty et al. [1], 77 other constituents were quantified according to the selective ion mode and two SPMEs were realized for each variety. So, chemical components were available for 16 strawberries $(S_i)_{1 \le i \le 16}$ and the dataset was expressed in the form of a matrix with 100 columns (chemical components) and 16 rows (analysed samples).

2.2 Pre-treatment

Chemical data were characterized by large differences between the magnitude levels of observed signals according to the measured variables. To standardise these data, obtained signals x relative to each selected aromatic constituent were transformed into y values according to the relation:

$$y = \log(x + 1) \quad (1)$$

and then converted into a centered reduced variable z as follows:

$$z = (y - \bar{y}) / \sigma_y \quad (2)$$

where $\bar{y}$ and σ_y are the mean and the standard deviation of the y value.

3 Modelling Procedure

3.1 SOM algorithm

Data were studied using the Kohonen self-organizing map (SOM) algorithm [8]. With this algorithm, the dataset will be projected in a non-linear way onto a rectangular grid laid out on a hexagonal lattice with N hexagons: the Kohonen map (Fig. 2). Formally the SOM consists of input and output layers connected with weight vectors. The array of neurones (i.e., computational units) in the input layer receives the input vectors (i.e. the standardized chemical components of the strawberries), whereas the output layer consists of a two-dimensional network of N neurons arranged on a hexagonal lattice.

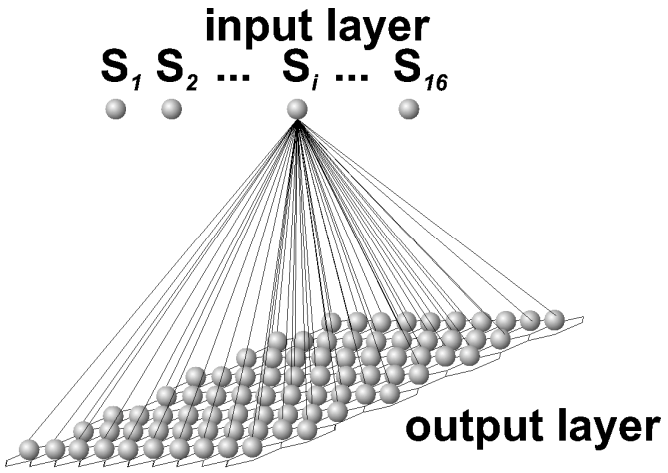

Fig. 2: Self-organizing map

For this purpose, in each hexagon, a reference vector will be considered. The reference vectors are in fact virtual strawberries $(VS_i)_{1 \le i \le N}$ with chemical constituents $(w_{ij})_{1 \le i \le N; 1 \le j \le 100}$ to be computed. In the output layer, the units of the grid (virtual strawberries) give a representation of the distribution of the strawberry samples in an ordered way. The modifications of the virtual strawberries are made through an Artificial Neural Network and computed during a training phase by iterative adjustments. For learning, only input units are used, no expected-output data is given to the system: this is referring to an unsupervised learning.

The standardized chemical compounds were used to train the SOM. In the learning process of the SOM, when an input vector (i.e., a real strawberry S_i) is sent through the network, the Euclidean distance between each virtual strawberry and the input vector is computed according to the chemical constituents. Among all N virtual strawberries, the best matching unit (BMU) characterised by the minimum distance between output and input vectors, is the winner. For the BMU and its neighbourhood neurons, the chemical constituents are iteratively updated in order to make them closer from the real strawberry. At the end of the training, the distribution of the virtual strawberries of the output layer reflects the distribution of the real strawberries. The detailed algorithm of the SOM can be found in Kohonen [8] for theoretical considerations, and Giraudel and Lek [9] and Urruty et al. [1] for ecological and chemical applications.

3.2 Chemical component distribution planes

When the learning process is finished, a map with S hexagons is obtained and in each hexagon, there is a virtual strawberry $(VS_i)_{1 \le i \le N}$ of which the chemical components $(w_{ij})_{1 \le i \le N; 1 \le j \le 100}$ have been computed.

During the learning process of the SOM, neurons that are topographically close in the array will activate each other to learn something from the same input vector. This approach results in a smoothing effect on the weight vectors of neurons [8]. Thus, these weight vectors tend to approximate the probability density function of the input vectors. Therefore, the visualization of elements of these vectors for different input variables is convenient to understand the contribution of each input variable with respect to the distribution of clusters in the SOM [9]. Calculated values (weights) of each input variable during the training process were visualized in each neuron of the trained SOM in grey scale. Therefore, the contribution of input variables (i.e., chemical components) appears clearly in the cluster structures defined in the SOM,.

Then, this map can be used in different ways:

- The sample strawberries can be mapped. For this purpose, the BMU is computed for each sample strawberry S_i and this one is put into the corresponding hexagon. The topological structure of the initial data set is preserved and similar fruits (for chemical composition) are displayed in the same hexagon or in neighbouring hexagons when dissimilar fruits are displayed in distant hexagons.
- The chemical constituents $(w_{ij})_{1 \le i \le N; 1 \le j \le 100}$ of each virtual strawberry can be used to display the distribution of each chemical compound selected. This representation can be considered as a "sliced" version of the SOM. Each plane displays each constituent in the virtual strawberries. For this purpose, a grey shade level was used: dark hexagons for high values of the constituent and light hexagons for low values of this constituent.
- The two former representations can be combined to display on the same figure the sample strawberries and the chemical constituents. By this way, several patterns can be observed: for instance:

- large regular gradients from one side of the map to an other one,
- small gradients for chemical components located only in an area of the map
- separated dark areas.

Each pattern corresponds to the main chemical traits of each fruit and by a visual way, the best chemical components for structuring of the map can be selected. But for large data sets, the observation of each map is a very hard work, so we have looked for an index able to give the structuring power of each chemical component: the Structuring Index (SI).

3.3 The Structuring Index

This Index was previously successfully used by the author, for ecological data [10]. The computation of the SI is made after the learning and is founded on the chemical components of the virtual strawberries. The SI was developed to define the most structuring chemical components in the distribution patterns of samples in the SOM. In other words, the SI is the value indicating relative importance of each chemical component in determining distribution patterns of samples in the SOM. The computation is made for each chemical component. So, for a component j, two elements have to be taken into account: firstly, the j^{th} component of each virtual strawberry but also the relative position of the virtual strawberries on the map. Formally, the contribution of each chemical constituent to the organisation of the map is depending of the absolute distance between the components: each couple of two VSs has to be taken into account. But, if the components of adjacent hexagons have to be similar, the distant hexagons may have more different components. So the distance between each couple of VSs is weighted by the inverse of the Euclidean distance *on the map* between the corresponding virtual units. So, for each chemical constituent, the SI is calculated from the summation of the ratios of the distance between the virtual strawberries according to this chemical constituent and the topological distance of two SOM units. It is expressed in the equation as follows:

$$SI_j = \sum_{k=1}^{N} \sum_{i=1}^{k-1} \frac{\left| w_{ij} - w_{kj} \right|}{\left\| r_i - r_k \right\|} \qquad (3)$$

where w_{ij} and w_{kj} are the values of the chemical constituent j in the SOM units i and k respectively, r_i and r_k are the coordinates of the units i and k respectively, and $||r_i - r_k||$ is the Euclidean distance between units i and k. N is the total number of the SOM output units.

The higher SI is, the better the contribution of the chemical component to the map structure.

Calculations have been performed with a PC equipped with an Intel® Pentium III-500 processor. The program file has been written by the authors (Giraudel and Lek 2001) using the version 1.5 of R software [11].

4 Results and Discussion

Some Kohonen maps of different sizes were trained. Increasing the map size, the process was stopped when all the strawberries appearing in the same unit were of the same variety. The discrimination of the 8 strawberry varieties was obtained using a 12-unit map (Fig. 3).

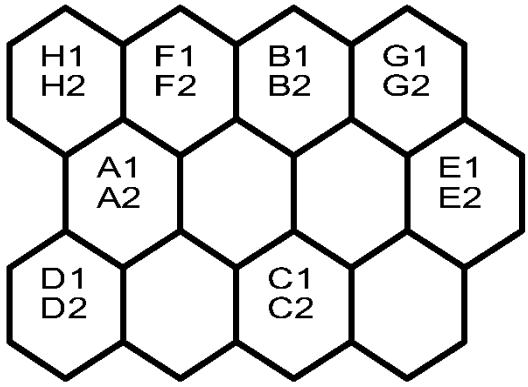

Fig. 3: Distribution of the samples in a 12-unit map built using all the chemical compounds.

Using the virtual strawberries of this map, component planes have been displayed for each chemical constituent. Figure 4 shows some examples of such a representation. Then, the Structuring Index has been computed for each chemical constituent. Firstly, the best constituent according to its SI can be considered:

Ethyl cinnamate (Fig. 4–a) separates the map into two parts: the upper left one with high concentration of this constituent and the lower right part with low concentration or absence of this constituent. Varieties H, F and A are clearly those with highest concentration. Ethyl cinnamate had the highest SI (SI=22.8).

This component shows a strong gradient of distribution in the map. Therefore, this component strongly contributed to determine patterns in the SOM.

Now, if the components with low SIs are considered:

- Heptanone-2 (Fig. 4-b) was a poor structuring component with only a very little area discriminated in the upper right area and it did not display specific characteristics in the major part of the SOM. Consequently, this component had a low SI (SI=14.6).
- 2propen-1-ol.3phenyl acetate (Fig. 4-c) had a non-regular distribution. It was abundant in varieties H and E located in two opposite corners of the map; so it did not mainly contribute to the organisation of the SOM (SI=17.1).

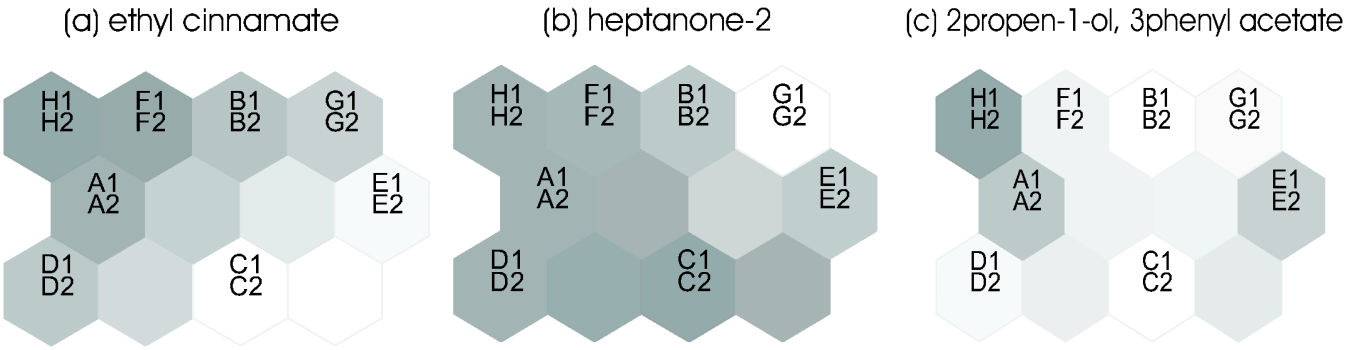

Fig. 4: Some component planes of the SOM

In order to validate the use of SI, a new 12 unit-SOM has been built, only using the 8 chemical constituents with highest SIs (Fig. 5).

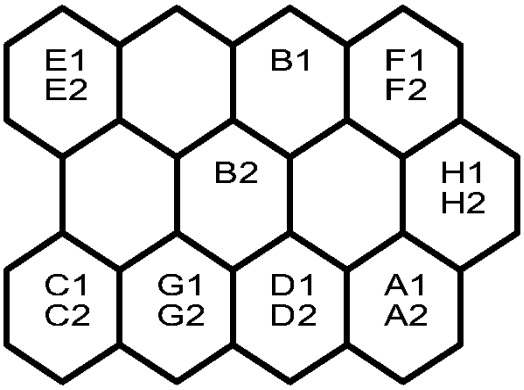

Fig. 5: Distribution of the samples in a 12-unit map built using the 8 chemical compounds with the highest SIs.

First of all, with this map, the complete discrimination was obtained: none hexagon contained two different varieties. The two samples of Ciflorette (B) were not located in the same hexagon but in adjacent ones. This new map was very consistent with the previously obtained map. The goal of the SOM is to display the topology of the dataset. Therefore, the orientation of the map (i.e. the position of the strawberries in one side of the map or in another one) is not important. So, the relative position of each variety remained perfectly consistent with the map obtained using all the chemical compounds. In this way, CF129 (A) and Capitola (H) et Cireine (F) remain neighbours in the left upper part of the map built using all the chemical compounds (Fig. 3) and in the right part of the map built using 8 chemical compounds (Fig. 5).

Among the 8 chemical compounds selected with the SI method, 3 were already used by Urruty et al. [1]: methyl butanoate, ethyl butanoate and butanoic acid, 5 were considered for the first time for strawberry variety aroma: ethyl cinnamate, hexyl isovalerate, benzyl alcohol and 2 other ones that remain to be chemically identified. Then, it appears that applying this SI method to a large dataset directly issued from the chromatogram, provide a promising approach and this point deserves to be more investigated using more strawberry varieties. Expected results should help to select the relevant constituents that should have to be analysed in respect with the desired objective (discrimination levels, chemical proximity of varieties,…). In the same way, the Structuring Index can be seen as a good tool for selecting the relevant chemical components in order to simplify the dataset before building supervised models. For instance, it could provide the reduction of the number of variables in backpropagation multilayer neural networks for prediction of sensorial analysis results.

References:

[1] L. Urruty, J.L. Giraudel, S. Lek, P. Roudeillac, M. Montury, Assessment of strawberry aroma through SPME/GC and ANN methods. Classification and discrimination of varieties. *Journal of Agricultural and Food Chemistry*. Vol.50, 2002, pp.3129-3136.

[2] R. Dutta, K.R. Kashwan, M. Bhuyan, E.L. Hines and J.W. Gardner, Electronic nose based tea quality standardization. *Neural Networks*, Vol.16, 2003, pp.847-853.

[3] P. Schieberle, and T. Hofman, Evaluation of the character impact odorants in fresh strawberry juice by quantitative measurements and sensory studies on model mixtures. *Journal of Agricultural and Food Chemistry*. Vol.45, 1997, pp.227-232.

[4] P. Grenier, Etat de l'art sur la technologie des nez électroniques. *Spectra analyse.* Vol.31, 2002, pp.30-35.

[5] J.L. Giraudel, and S. Lek, Ecological applications of unsupervised artificial neural networks. In, *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation* (F. Recknagel, Ed.) Springer, Berlin, 2003, pp.15-32.

[6] A. Latrasse, Fruits III. In, *Volatile Compounds in Foods and Beverages* (Maarse, H. Ed.) Dekker, New York, 1991, pp.333-387.

[7] J. Pawliszyn, *Solid Phase Micro Extraction : Theory and Practice*; Wiley-VCH: New-York, 1997.

[8] T. Kohonen, *Self-Organizing Maps*. Third edition, Springer, Berlin, 2001

[9] J.L. Giraudel and S. Lek, A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling*. Vol.146, 2001, pp.329-339.

[10] J. Tison, J.L. Giraudel, M. Coste, F. Delmas, Y.S. Park, Use of unsupervised neural networks for ecoregional zoning of hydrosystems through diatom communities: case study of Adour-Garonne watershed (France). *Archiv fur Hydrobiologie*. (accepted)

[11] R. Ihaka, R. Gentleman, R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. Vol.5, 1996, pp.299-314.