Journal of Mathematical Chemistry, Vol. 40, No. 1, July 2006 (© 2006)
DOI: 10.1007/s10910-006-9118-5

A method for clustering and screening of long-dimensional chemical data based on fingerprints and similarity measurements

Manuel Urbano Cuadrado, Gonzalo Cerruela García, Irene Luque Ruiz,* and Miguel Ángel Gómez-Nieto

Department of Computing and Numerical Analysis, University of Córdoba, Campus Universitario de Rabanales, Albert Einstein Building, E-14071 Córdoba, Spain
E-mail: ma1lurui@uco.es.

A method for the treatment of long-dimensional chemical data arrays is presented in this work with the aim of maximising classification models. The method is based on the construction of fingerprints and the subsequent generation of a similarity matrix. The similarity calculation has been modified through a scaling process to take into account different significance shown by the variables. The method was applied to spectral measurements of wines and several aspects were studied, namely: threshold considered in the construction of fingerprints and patterns, weighting factor used for scaling, normalisation method, etc. The application of both Principal Components Analysis and Soft-Independent Modelling of Class Analogies to the similarity matrices gave better classifications of the information than those obtained using original data.

KEY WORDS: data preparation, similarity calculation, fingerprints, clustering, screening

MSC 2000: 68T10, 62H30, 93C35

1. Introduction

Data employed for modelling of natural or artificial processes can be obtained from scientific and engineering experiments. Modern instrumental techniques provide scientists with Long-Dimensional Data Arrays (LDDA) in short intervals of time. The information able of extracting from these LDDAs depends considerably on the applicability of mathematical and statistical methods to these data sets. Multivariate analysis is the statistical discipline that encompasses methods dealing with the study of phenomena or objects characterised by n observations or properties, respectively [1,2].

Supervised Pattern Recognition Techniques (SPRT). The assignment of an object O_i into a given class C_j can be expressed as a function of a set of

*Corresponding author.

variables $V = \{v_1, v_2, \ldots, v_m\}$. If the measurement of V is not easy (expensive, time-consuming, etc.), a set of predictor variables $P = \{p_1, p_2, \ldots, p_j\}$ with higher availability than V can be used to predict when an object belongs to a given class. This is considered as a qualitative regression process consisting of two steps, namely: the training and testing stages. This classification process is indirect, and so, SPRTs [3,4] must be used in order to build the classification model.

There are two sorts of SPRTs summarised in literature, namely: parametric and non-parametric approaches. The formers are based on the calculation of distances between the different objects. Mahalanobis and Euclidean distances [5,6] are the two main parameters employed for classification. In chemistry, the majority of parametric classification methods are based on Soft-Independent Modelling of Class Analogies (SIMCA) and Linear Discriminant Analysis (LDA) techniques. A reduction of the original data space into a latent space is carried out in both above commented techniques.

Regarding non-parametric approaches, the use of Artificial Neural Networks (ANN) and Genetic Algorithms (GA) with the aim of classifying objects has been extensively used in science. Here, the assignment of a given object to a specific class is based on concepts related to the functioning of the human brain and the evolution theory [7].

The testing step consists of the evaluation of the prediction capability of the developed classification rules. The classification error (considered as the sum of false positives and false negatives obtained for the test set) is the parameter employed for the characterisation of the models' efficiency [8]. This can be negatively affected by two main reasons, namely: low signal/noise ratios owing to the high-random error involved in some types of measurement and LDDAs showing a high degree of similarity between them [9,10].

In this work, a new method for transforming LDDAs into fingerprints and the subsequent construction of a similarity matrix is presented with the aim of overcoming the upon above-commented disadvantages. After describing the method, its efficiency as data treatment in the development of classification rules was tested. With this purpose, SIMCA models were built using similarity matrixes as inputs of SIMCA processes.

2. Translating data to fingerprints

2.1. Outliers detection and deletion

Detection of LDDAs that behave as outliers is carried out to detect anomalies. Objects are projected in the space built after realising a PCA, and then, leverage value (a measure of how far an object is compared to the majority) and

residual variance are computed. The outliers must be examined in order to know if either they provide any useful information or they have to be removed.

2.2. Data normalisation

A LDDA is considered as a variable array a composed by n elements $a(i)$ that represent the measurements at different conditions i. A matrix A with dimensions $n \times a$ is defined for the sample set, where n is the number of samples and a is the number of variables. The element $A(i, j)$ represents the measurement value for the sample i at the condition j. The matrix A is transformed into a normalised matrix $\bar{A}$.

In this paper, the standard normalisation has been used, expressed as follows:

$$\text{Standard}: \ \forall n, \overline{A(i, j)} = \frac{A(i, j) - \min(A(n, j))}{\max(A(n, j)) - \min(A(n, j))}. \tag{1}$$

Other types of normalisation were tested. The types depending on data distribution (as the standard normalisation), i.e. logarithmic normalisation, yielded similar results. On the other hand, normalisation methods independent of data distribution, i.e. tangential normalisation, were not appropriate.

2.3. Construction of the fingerprints

After normalising data, the matrix A has been transformed into a new matrix $\bar{A}$ consisting of values within the range [0,1]. The latter is used for building a new matrix F, the fingerprint matrix, as follows:

- A threshold value t within the range [0,1] is selected for the construction of fingerprints.
- An element $F(i, j) = 1$, if and only if $\overline{A(i, j)} \geqslant t$, and $F(i, j) = 0$ otherwise.

Thus, the threshold value t determines the significance of the measurement at every condition. As an extreme case, the matrix F is equal to the unity matrix when $t = 0$ and, on the other hand, most of elements $F(i, j)$ are equal to zero when $t = 1$.

3. Obtaining similarity matrices

Once fingerprints have been generated, a symmetrical matrix of similarity is built with a low computational cost. The transformation of the LDAs into fingerprints allows using any of the similarity indexes proposed in the literature

[11,12]. In this paper, the Tanimoto index has been used, which is described as follows:

$$T_{A,B} = \frac{c}{a + b - c}, \tag{2}$$

where c represents the number of bits set to 1 common in the fingerprints A and B; a represents the number of bits set to 1 in the fingerprint A; and b represents the number of bits set to 1 in the fingerprint B.

A new way of similarity calculation based on the Tanimoto index is proposed aimed at increasing the robustness of the methods regarding internal variability of the samples. Thus, the averaged Tanimoto index takes into account the correspondence of the bits equal to 1 ($T^1_{A,B}$) and the bits equal to 0 ($T^0_{A,B}$) between two fingerprints. The averaged index is calculated as follows:

$$\overline{T_{A,B}} = \frac{T^1_{A,B} + T^0_{A,B}}{2}. \tag{3}$$

The calculation of similarity measurements using the Tanimoto index is applied to the matrix F, thus building the similarity matrix S. This matrix is symmetrical and its dimension is $n \times n$, where n is the number of samples (objects). The elements $S(i, i)$ are equal to 1 and each element $S(i, j)$ represents the similarity value between the sample i and the sample j obtained from the application of equations (2) or (3) to the fingerprint matrix F.

4. Refining the similarity matrix calculation

As can be seen in equations (2) and (3), the following points have been considered in the construction of similarity matrices:

1. All the bits of the fingerprints, and in turn, all the condition measurements, contribute to the calculation with a constant level of significance. Thus, all the bits influence the characterisation and behaviour of the samples in the same way. This influence or loading is equal to $1/n$, where n is the total number of variables.
2. Therefore, all the samples belong to a set with similar characteristics. This set is used for comparison by means of a similarity calculation.

Since the target objective is the development of methods for classification of objects, the calculation of the similarity matrix S can be enhanced. The procedure proposed is as follows.

4.1. Construction of pattern fingerprint

The number and position of the bits set to 1 (or 0) in the pattern fingerprint are the keys for the application of equations (2) or (3), and they clearly depend

on the problem under study. For that, this dependence makes necessary that the proposed model is open regarding any problem, any number of variables, characteristics of the objects, etc.

The existence of a set of pattern objects with properties very well-known possibilities the construction of a F^P matrix composed by p rows (number of pattern objects) and a columns or variables (equal to the F matrix).

The frequency of 1s is analysed for each column a of the F^P matrix in order to generate an array of frequencies f in the following way:

$$f(a) = \frac{\sum_{i=1}^{i=z} F^P(i, a)}{p}, \tag{4}$$

where the values i determine the samples to be considered for construction of the pattern, z is the number of bits set to 1 and p is the number of pattern samples.

Now, a pattern fingerprint P^1 is built from the frequency array f by the consideration of a threshold frequency value t'. Thus, a pattern fingerprint P^1 is generated with elements $P^1(i) = 1$ if the corresponding element of the frequency array $f(i)$ has a value equal to or higher than the threshold value t'. As an extreme case, all the bits of the pattern fingerprint P^1 are set to 1 when the threshold value is $t' = 0$. Thus, the higher the values for t', the lower the number of bits set to 1 in the pattern fingerprint.

4.2. *Scaling the similarity matrices*

As can be seen in equations (2) and (3), all the fingerprint bits have the same influence on the similarity calculation. This assertion is correct in applications as structural similarity calculation [13]; nevertheless, considering a constant value for the influence of the bits is incorrect in other many problems [14, 15]. Thus, cases like data acquired at very different conditions, objects with different characteristics, etc., make necessary to consider different loadings for the fingerprint bits. Also, obtaining clusters is very difficult when the objects are very similar owing to the few differences between the fingerprints.

Therefore, the similarity calculation can be modified in order to take into account different loadings for the bits. This modification is carried out through a scaling process using a pattern fingerprints P^1 of dimension equal to those of the fingerprints that form the matrix F. Bits set to 1 in the pattern represent the group of bits more significant of the bits set to 1 in the F matrix, that is, the conditions that provide more information from the object under study.

Thus, a new calculation of the similarity matrix is proposed by the application of a scaled Tanimoto index as follows:

$$T_s = \frac{c_s}{a_s + b_s - c_s}, \tag{5}$$

where

$$\begin{aligned} a_s &= a_a + a_p \; x \; W, \\ b_s &= b_b + b_p \; x \; W, \\ c_s &= c_c + c_p \; x \; W, \end{aligned} \tag{6}$$

- The a_a, b_b and c_c values represent the number of bits set to 1 in the fingerprints A, B and common in A and B, respectively, and not set to 1 in the pattern fingerprint P.
- The a_p, b_p and c_p values represent the number of bits set to 1 in the fingerprints A, B and common in A and B respectively, and set to 1 in the pattern fingerprint P.
- W is a scaling factor that permits to give more weight to those bits set to 1 in the samples A and B and also set to 1 in the pattern P^1 in the similarity calculation.

As can be seen, equation (6) permits to weight up the different fingerprint bits in the similarity calculation, thus considering different levels of significance for the variables.

A pattern fingerprint P^0 can be built considering the frequency of the bits set to 0 and the threshold t'. Both pattern fingerprints (P^1 and P^0) are used when the averaged Tanimoto index is employed for the generation of the similarity matrix.

5. Classifying information: soft independent modelling of class analogy

After building the similarity space, this is used as the input of classification processes. The SIMCA is a technique employed for the development of rules capable of determining if new objects belong to an already existing group. The use of SIMCA in contrast to other pattern recognition approaches is based on the versatility involved in SIMCA rules. Thus, SIMCA allows assigning a target object to: (a) none of the modelled classes; (b) only one class; and (c) two or more classes of those modelling the similarity space considered.

5.1. Developing the classification rules: the learning stage

The development of SIMCA models involves two steps, namely: the learning and testing stages. In the first step, objects with class-keyed properties known are used to build the membership space from predictors (in this case, similarity values). For this, a *Principal Component Analysis* (PCA) is realised for each group of learning objects belonging to a given class. The aim of this step is

the transformation of original data to a reduced space for establishing of the membership class zone in a way easier than that for the original space. This is due to removal of co-linearity in original data and, in turn, the use of few latent variables that are linear combinations of original similarity values.

The theorem of matrix algebra, called *Singular Value Decomposition* (SVD), is used for the upon above-commented transformation as follows:

$$S_C = U \Lambda V^T, \tag{7}$$

$$S_C = \lambda_1 u_1 V_1^T + \lambda_2 u_2 V_2^T + \cdots + \lambda_r u_r V_r^T, \tag{8}$$

where S_C is the $n \times p$ similarity matrix (n is the number of objects that compose the learning set of a class C and p is the number of similarity variables); U is an nxr column-orthonormal matrix; V is a pxr column-orthonormal matrix; and Λ is an rxr diagonal matrix. Equation (8) is the compact equation (7) written as individual contributions. The fact of the orthonormality required for the new space implies that:

$$U^T U = P^T P = I_r. \tag{9}$$

Traditionally, the notation employed in chemometrics for SVD defines the score and loading matrixes (after carrying out a series of simplifications, whose explanation is out of the scope of this work), represented by means of the symbols T and P such that $T = U\Lambda$ and $P = V$. Thus, equations (7) and (8) are transformed as follows:

$$S_C = T P^T, \tag{10}$$

$$S_C = t_1 p_1^T + t_2 p_2^T + \cdots + t_r p_r^T. \tag{11}$$

In addition to orthonormality requirements, the construction of the principal components is based on the criterion of maximal variance of data explained by these factors or latent variables. Thus, SVD is carried out keeping the subspace with largest variance. The first principal component is selected as follows:

$$p_1 = \max\{\mathrm{var}(P^T S_C)\}. \tag{12}$$

For the $k-1$ component, the following calculation is carried out:

$$\widehat{S}_{C_{k-1}} = X - \sum_{i=1}^{k-1} p_i p_i^T X, \tag{13}$$

$$p_k = \max\{\mathrm{var}(P^T \widehat{S}_{C_{k-1}})\}. \tag{14}$$

Since only the most significant latent vectors are retained in practical situations, equation (10) is transformed in the following equation:

$$S_C = T P^T + N, \tag{15}$$

where N is the residual data matrix. The optimal number of scores (r') needed to describe the structure of the learning class C can be determined by cross-validation. The residuals from the model can be computed from the scores on the non-retained eigenvectors. Then:

$$\text{res} = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=r'+1}^{r} t_{i,j}^2}{[(r - r')(n - r' - 1)]}}. \tag{16}$$

5.2. *Validating the classification rules: the testing stage*

After developing the model for the class C, its prediction efficiency must be tested with an independent set of objects (the testing set). The statistical parameter used is the classification error, which is computed as the sum of false positives (C non-member objects assigned to the class C) and false negatives (C member objects not assigned to the class C) divided by the number of testing objects. The evaluation of this error is depending on the problem (target degree of classification, time reduction achieved, number of objects with information available, etc.).

The criterion for assigning a new object to the class C is based on a critical value of Euclidean distance towards the model. This is given by:

$$\text{res}_{\text{crit}} = \sqrt{F_{\text{crit}}}\,\text{res}. \tag{17}$$

For a new object, the Euclidean distance from the model is then obtained, similarly to equation (16):

$$\text{res}_{\text{new}} = \sqrt{\frac{\sum_{j=r'+1}^{r} t_{\text{new},j}^2}{(r - r')}}. \tag{18}$$

Finally, if $\text{res}_{\text{new}} < \text{res}_{\text{crit}}$, the new object belongs to class C, otherwise it does not.

6. Application of the method to spectroscopy data

The proposed method has been applied to spectral data obtained by means of the MIR technique. Each array consisted of 1142 variables and different types of wines were studied.

A high degree of similarity between the wine samples can be observed in figure 1. The exception for this similarity is the 1150–1300 and 2300–2400 cm^{-1} zones. The use of the proposed method for enlarging of the dissimilarities is necessary in order to differentiate and classify samples of wines using spectroscopic data under various criteria.

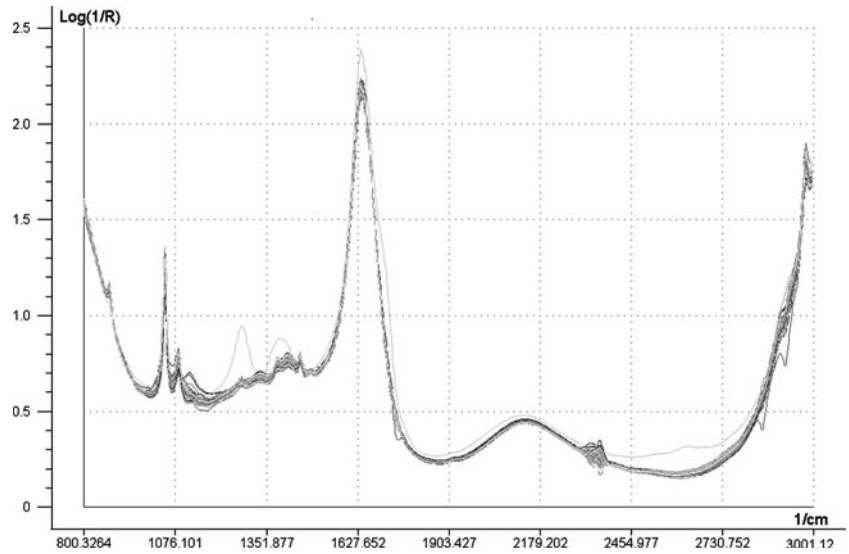

Figure 1. MIR spectroscopic data for all the samples of wine.

Analysis of spectral outliers was carried out to avoid the influence of anomalous spectra on building of the fingerprints, and then, on the generation of the similarity matrices.

Spectra were normalised using the standard method (equation 1). A normalised spectrum of a sample selected randomly and the fingerprints built using different threshold values (t) are shown in figure 2.

As figure 2 shows, the selection of the threshold value t is a key aspect in building of fingerprint as it determines its density. An increasing of threshold value produces a decreasing of fingerprint density. Fingerprints with high-density yielded high-similarity values, even for very different samples. Just the opposite occurs with low-density fingerprints. Thus, very low-similarity values were obtained, even for very similar samples.

A comparison between the use of similarity matrices (using Tanimoto index and averaged Tanimoto index) and the spectral data is given in figure 3. In this figure, the score plots for the PCA applied to both the similarity matrices generated from fingerprints and the spectral data matrix are shown. Different threshold values t (from 0 to 1) were used for building of fingerprints. Extreme threshold values produced an anomalous behaviour that yielded a high no-explained data variance. The best discrimination was achieved for the threshold value $t = 0.4$ and the averaged Tanimoto index, as figure 3 shows.

The proposed method permits to increase and decrease the similarity values for similar and different samples, respectively. Two factors influence on the efficiency of the pattern for enlarging of the differences: (a) the selection of

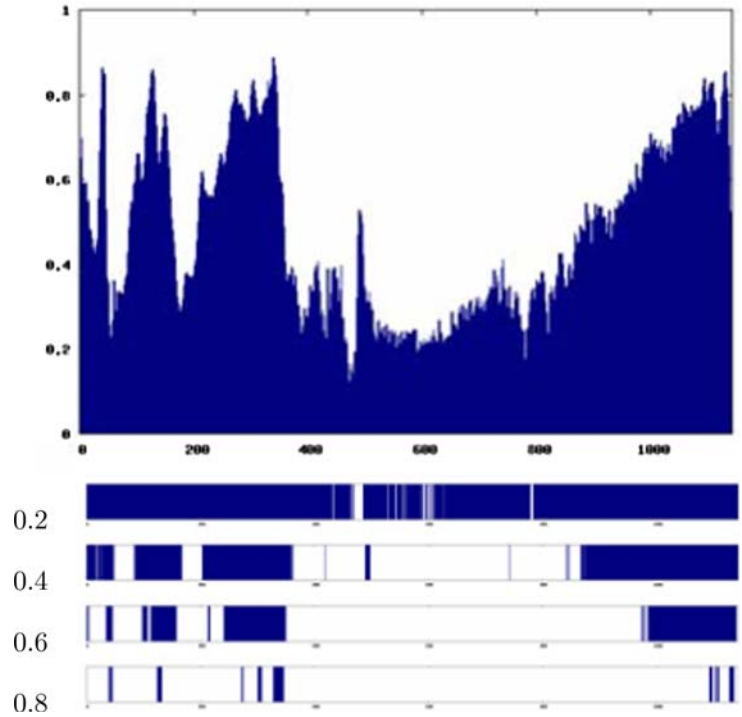

Figure 2. Normalised spectrum of an object selected in a random way and the fingerprints built using different threshold values.

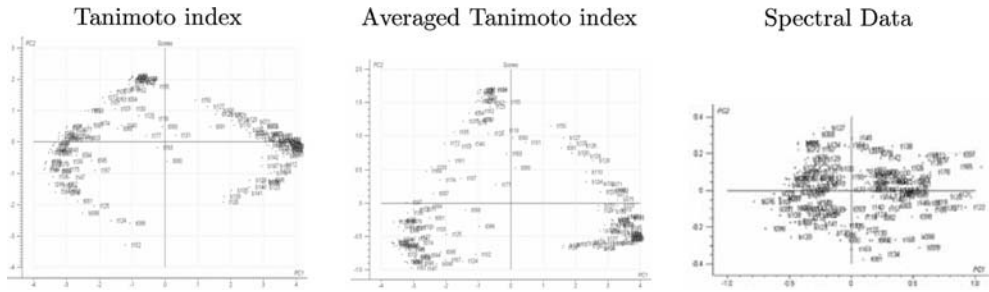

Figure 3. Score plots resulting from applying PCA to the similarity matrices (Tanimoto and averaged Tanimoto indexes using a threshold value $t = 0.4$) and the spectroscopic data matrix. *b*: white wines, *t*: red wines.

the samples used for building of the pattern fingerprint; and (b) the frequency threshold value t' employed.

Figure 4 shows both the frequency distribution of the normalised values and the corresponding pattern fingerprints generated from different threshold values t' for Cencibel variety aged and young wines. Thus, the proposed method

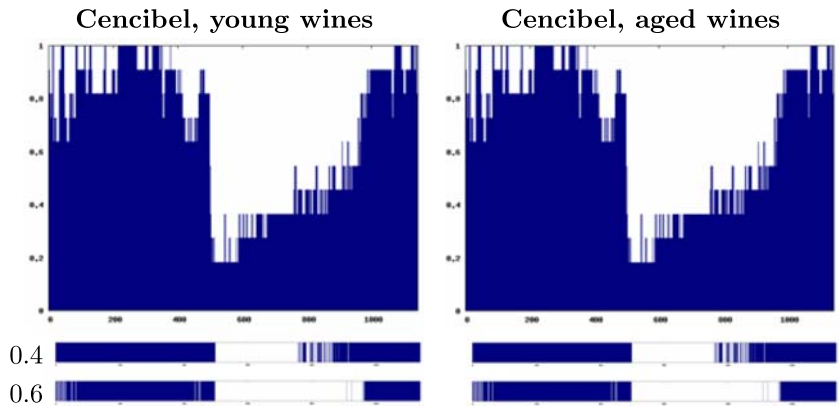

Figure 4. Frequency distributions and their pattern fingerprints for $t' = 0.4$ and $t' = 0.6$ using: Cencibel variety aged and young wines.

based on building of fingerprints from spectral information permits to differentiate wines.

Low-threshold values t' yielded pattern fingerprints with high density that were not appropriate for samples classification. Nevertheless, the fingerprints generated using a threshold value $t' = 0.6$ were appropriate for characterisation.

7. Discussion and remarks

A new method based on use of fingerprints and scaling technique for the similarity calculation has been proposed in this paper for the development of classification models. The method has been applied in an efficient way to spectral data of wines of different type, origin and grape variety.

The capacity for classification of the different models has been validated in a series of tests using SIMCA applied to similarity matrices. A comparison between the similarity matrices and the spectral matrix was carried out. Thus, a training set composed by 85 samples and a validation set consisting of 30 samples (not considered for the training set) were selected.

The error obtained in the classification of white and red wines was 20% for spectral data. When the similarity matrices were employed the error decreased to 5% with the use of the threshold value $t = 0.4$, pattern fingerprints of $t' = 0.6$, and $W = 9$. The similarity matrices were generated using the averaged Tanimoto index.

Efficiency of the models for classifying wines according to grape variety, origin and ageing process was also improved. Although the transformation of original data to similarity values involves information removal, the part of the signal removed is that stochastic or random part, which hampers the interpretation of the deterministic part.

Table 1
Use of different patterns for similarity calculation between two wines from Cencible grape and two wines from Cabernet Sauvignon grape. These wines correspond to two different cultivation areas (Origins 1 and 2). The wines have been randomly selected.

	Cencibel-1	Cencibel-2	Cabernet-1	Cabernet-2
Without Pattern				
Cencibel-1	1.00	0.76	0.56	0.52
Cencibel-2		1.00	0.51	0.60
Cabernet-1			1.00	0.73
Cabernet-2				1.00
Cencilbel Pattern				
Cencibel-1	1.00	0.96	0.51	0.49
Cencibel-2		1.00	0.46	0.56
Cabernet-1			1.00	0.72
Cabernet-2				1.00
Cabernet Pattern				
Cencibel-1	1.00	0.75	0.54	0.52
Cencibel-2		1.00	0.48	0.56
Cabernet-1			1.00	0.91
Cabernet-2				1.00
Origin-1 Pattern				
Cencibel-1	1.00	0.72	0.77	0.50
Cencibel-2		1.00	0.50	0.53
Cabernet-1			1.00	0.73
Cabernet-2				1.00
Origin-2 Pattern				
Cencibel-1	1.00	0.75	0.48	0.50
Cencibel-2		1.00	0.51	0.81
Cabernet-1			1.00	0.70
Cabernet-2				1.00

An example of the effect produced by the use of different patterns in the calculation of the similarity is shown in table 1. Samples selected in a random way corresponding to two types of wines (Cencibel and Cabernet–Sauvignon varieties) and two different origin zones (Origins 1 and 2) were used. As can be observed in table 1, the use of patterns increases the similarity between the wines belonging to the pattern, in addition to the decreasing of the similarity for different types. Moreover, the use of patterns with a high degree of specificity (grape variety) produces changes more significant than those obtained when the patterns are less specific (origin denominations). In spite of this, an improvement in the samples discrimination is observed in all the cases.

Pattern fingerprints databases with low requirements for storage can be built and updated with new data in order to refine pattern construction.

Acknowledgments

The Comisión Interministerial de Ciencia y Tecnología (CICyT) is thanked for financial support (Project TIN2004-04114-C02-01).

References

[1] K. H. Esbensen, *Multivariate Data Analysis – in Practice* (Camo Process AS, Norway 2002).
[2] O.F. Alis and R. Herschel, J. Math. Chem. 29 (2001) 127–142.
[3] G. Hagberg, NMR Biomed. 11 (1998) 148–156.
[4] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition* (Wiley, New York, 1992).
[5] R. DE Maesschalck, D. Jouan-Rimbaud and D.L. Massart, Chem. Intell. Lab. Syst. 50 (2000) 1–18.
[6] P.C. Mahalanobis, Proc. Nat. Inst. Sci. India 12 (1936) 49–55.
[7] R. Leardi, *Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks* (Elsevier, Amsterdam, 2003).
[8] E. Trullos, I. Ruisanchez and F. X. Rius, Trends Anal. Chem. 23 (2004) 137–145.
[9] B. Walczak and D. L. Massart, Chem. Intell. Lab. Syst. 36 (1997) 81–94.
[10] J. Steiner, Y. Termonia and J. Deltour, Anal. Chem. 44 (1972) 1906–1909.
[11] P. Willet, J. M. Barnard and G. J. Downs, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[12] D. H. Rouvray and A. T. Balaban, In: *Chemical Applications of Graph Theory. Applications of Graph Theory*, eds. R.J. Wilson and L. W. Beineke (Academic Press, London, 1979).
[13] K. Varmuza, M. Karlovits and W. Demuth, Anal. Chim. Acta 490 (2003) 313–324.
[14] L. Xue, F. L. Stahura, J. W. Godden and J. Bajorath, J. Chem. Inf. Comput. Sci. 41 (2001) 746–753.
[15] P. Mazzatorta, E. Benfenati, D. Neagu and G. Gini, J. Chem. Inf. Comput. Sci. 42 (2002) 1250–1255.