Pattern Analysis and Applications

10.1007/s10044-010-0187-5

Industrial and Commercial Application

A three-step unsupervised neural model for visualizing high complex dimensional spectroscopic data sets

Emilio Corchado [1] and Juan C. Perez [2]

(1) Departamento de Informática y Automática, University of Salamanca, Salamanca, Spain
(2) Department of Civil Engineering, University of Burgos, Burgos, Spain

Emilio Corchado (Corresponding author)
Email: escorchado@usal.es

Juan C. Perez
Email: jcperez@ubu.es

Received: 12 November 2009
Accepted: 22 August 2010
Published online: 12 October 2010

Abstract

The interdisciplinary research presented in this study is based on a novel approach to clustering tasks and the visualization of the internal structure of high-dimensional data sets. Following normalization, a pre-processing step performs dimensionality reduction on a high-dimensional data set, using an unsupervised neural architecture known as cooperative maximum likelihood Hebbian learning (CMLHL), which is characterized by its capability to preserve a degree of global ordering in the data. Subsequently, the self organising-map (SOM) is applied, as a topology-preserving architecture used for two-dimensional visualization of the internal structure of such data sets. This research studies the joint performance of these two neural models and their capability to preserve some global ordering. Their effectiveness is demonstrated through a case of study on a real-life high complex dimensional spectroscopic data set characterized by its lack of reproducibility. The data under analysis are taken from an X-ray spectroscopic analysis of a rose window in a famous ancient Gothic Spanish cathedral. The main aim of this study is to classify each

sample by its date and place of origin, so as to facilitate the restoration of these and other historical stained glass windows. Thus, having ascertained the sample's chemical composition and degree of conservation, this technique contributes to identifying different areas and periods in which the stained glass panels were produced. The combined method proposed in this study is compared with a classical statistical model that uses principal component analysis (PCA) as a pre-processing step, and with some other unsupervised models such as maximum likelihood Hebbian learning (MLHL) and the application of the SOM without a pre-processing step. In the final case, a comparison of the convergence processes was performed to examine the efficacy of the CMLHL/SOM combined model.

Keywords Unsupervised learning – Projection methods – Topology-preserving mapping – Visualization – Spectroscopic analysis

---

1 Introduction

Unsupervised learning is an important facet of the human learning experience and is central to many artificial neural networks (ANNs). A wide variety of unsupervised learning architectures have been proposed to date. The novel pre-processing step presented in this study, based on an exploratory projection pursuit (EPP) method [*15*, *16*], is applied to overcome two of the main disadvantages of the self-organising maps (SOM) [*2*]: low convergence speeds and difficulties achieving convergence in complex domains.

The effectiveness of this approach is demonstrated through a case study of a complex high-dimensional spectroscopic data set. Electronic microscopy is one of the most widespread techniques in materials analysis. Macrophotography and X-ray emission analysis of the sample, respectively, provide information on the sample's morphology and composition.

The spectroscopic technique is used to assess the concentration or amount of a given chemical (atomic, molecular, or ionic) species. The technique is based on a scattering process, in which crystalline materials scatter X-rays at well-defined angles. If the wavelength of the incident X-rays is known, the distances between planes of atoms within the crystal may be calculated. The intensities of the scattered X-rays provide information on the atomic positions and allow the atomic distribution in the crystal structure to be calculated [*18*, *19*].

The application of these spectroscopic techniques to the characterization of historical components using minimum amounts of sampling material, and with no physical contact between the instrument and equipment, are non-destructive analytic techniques that respect the integrity of the material under study.

The results of the X-ray emission spectrograph are qualitative and semi-quantitative in

nature. This method generates a great quantity of information, which has to be easily and efficiently identified and classified.

The internal structures of complex clustering domains, such as high-dimensional spectroscopic ones, can sometimes hinder identification of their patterns. These patterns may become visible if a change is made to the basis of the space; however, as a priori decision as to which basis will reveal most patterns requires an ability to predict unknown patterns. Furthermore, some variables may be redundant if the information they add is contained in other variables. Others may contain false correlations which can complicate the process of detecting the underlying patterns in the data. Extra variables may also increase computation time and can interfere with the accuracy of the clustering or classification process.

Feature selection improves classification by searching for the subset of features from the original set of variables which best classifies the training data [*1*]. Feature selection and extraction include feature construction, space dimensionality reduction, and sparse representations, among others. All these techniques are commonly used as pre-processing tools in pattern recognition [*42*] and other machine learning tasks. Although researchers have had to grapple with such problems for many years, interest in feature extraction has recently been renewed. Space dimensionality reduction is critical for the efficiency and efficacy of the predictors in a large number of new applications with very large input spaces [*38*]. Some of these applications include new and classical topics such as bioinformatics (DNA microarrays, etc.) [*39–41*], remote sensing multi and hyperspectral imagery, pattern recognition [*42*], (e.g. handwriting recognition, text processing, web mining) [*43*, *44*], speech processing [*48*], artificial vision [*47*, *49*], intrusion detection [*45*] and so on.

Our approach to feature selection is based on space dimensionality reduction. It initially uses a projection method called cooperative maximum likelihood Hebbian learning (CMLHL) [*10*], characterized by its capability to enforce a sparser representation in each weight vector than other classical methods such as PCA [*11*, *12*] or maximum likelihood Hebbian learning (MLHL) [*6*]. Of more importance in this study is its capability to preserve some global ordering in the data set, due to the effect of lateral connections. This is a very interesting approach to reproducibility problems presented by the data set under analysis, as will be explained in Sect. 5. CMLHL yielded successful results when initially applied to the analysis of several internal data set structures [*28–30*, *36*].

Having significantly decreased data dimensionality, the SOM [*2*] is applied to the pre-processed data set for visualization purposes. This ANN was chosen due to its topological preserving capability, which is very useful when confronted with the poor reproducibility of the problem presented in this research.

The rest of the paper is structured as follows: Sect. 2 presents a short overview of the unsupervised learning method and the well-known unsupervised model: self-organising

maps. Section 3 introduces some projection techniques as PCA, MLHL and CMLHL, which are the pre-processing methods applied and compared in this research. Section 4 describes an interesting real-life complex high-dimensional spectroscopic data set. Section 5 describes the unsupervised connectionist combined method in detail. Section 6 goes on to present the experiments and results. A comparison with other methods is presented in Sect. 7, which details the convergence process of the combined method in comparison with the standard SOM. Finally, the conclusions and future research work are presented in the final section.

---

2 Unsupervised learning

Human beings appear to be able to learn without explicit supervision. Unsupervised learning attempts to use methods that mimic biological processes associated with human learning in a more plausible way than error descent. For example, unsupervised learning algorithms have local processing at each synapse, avoiding the need for global information passing. So, an unsupervised neural network must self-organise with respect to its internal parameters, without external prompting, and to do so, it must react to some aspect of the input data. Typically, this will be either redundancy in the input data or clusters in the data; i.e. there must be some structure in the data to which it can respond.

All the artificial neural networks (ANNs) used in this research are based on unsupervised learning. The most representative unsupervised bio-inspired model is the SOM or Kohonen map [*2*, *9*].

2.1 Kohonen map

Kohonen [*2*] developed the self-organising map (SOM) as a visualization tool for high-dimensional data on a low-dimensional display. A SOM is composed of a discrete array of L nodes arranged on an N-dimensional lattice and it maps these nodes into a D-dimensional data space while maintaining their ordering. The SOM has been successfully applied to a wide variety of applications. It has found wide applications in text mining [*20–22*], data mining [*23*], web mining [*37*], process analysis [*24*], biological signal analysis [*25*], images [*8*], image coding [*26*] and so on.

The literature on SOM applications appears to lack a thorough theoretical analysis of certain SOM-related issues, such as topology preserving, convergence analysis, performance analysis, and in particular, multidimensional inputs [*3–5*, *27*]. One of the SOM's problems is its low convergence speed. Thus, this study sets out a way of speeding up the convergence process. It is based on the use of a fast pre-processing method characterized by its capability to enforce a sparser representation in each weight vector and more importantly for this study,

by its capability to preserve some global ordering [*10*]. We consider it an appropriate pre-processing method for complex high-dimensional data sets, which may be used in combination with topology-preserving methods such as the SOM, due to its feature preservation of global ordering.

---

3 Projection methods

3.1 Principal component analysis

Principal component analysis (PCA) [*11*, *12*] is a statistical technique which seeks to find the orthogonal basis that maximizes the projection variance for a given dimensionality of basis. This usually involves finding the direction which accounts for most of the data variance, which becomes the first principal component. The next component is the direction from the remaining data which contains the most variance and is orthogonal to the previous basis vector.

PCA can also be thought of as a data compression technique which involves minimum information loss in the data, in terms of least mean squared error. As a result, it is often used as a pre-processing method in order to simplify further analysis.

Taking an analysis by [*33*], it is possible to describe this as mapping vectors $\mathbf{x}^{\mathbf{d}}$ in an N-dimensional space $(x_1, \ldots, x_N)$ onto vectors $\mathbf{y}^{\mathbf{d}}$ in an M-dimensional space $(y_1, \ldots, y_M)$, where $M \leq N$, $\mathbf{x}$ may be represented as a linear combination of a set of N orthonormal vectors W_i.

PCA can be also implemented by means of several connectionist architectures [*13*, *14*, *17*].

3.2 A cooperative neural pre-processing method

The pre-processing architecture used in this research is based on an exploratory projectionist model (EPP) [*15*, *16*] called CMLHL [*10*], which was initially applied in the field of Artificial Vision. It is based on another EPP method called MLHL [*6*]. The application of lateral connections derived from the rectified Gaussian distribution (RGD) [*7*] to enforce a sparser representation of each weight vector means that it is capable of preserving some global ordering. MLHL is based on a family of cost functions which maximizes the likelihood of identifying a specific distribution.

If there is an N-dimensional input vector, x, and a M-dimensional output vector, y, with W_{ij} being the weight linking input j to output i, then MLHL can be expressed as

$$y_i = \sum_{j=1}^{N} W_{ij} \cdot x_j, \forall i \tag{1}$$

The activation (e_j) is feedback through the same weights and subtracted from the input.

$$e_j = x_j - \sum_{i=1}^{M} W_{ij} \cdot y_i, \forall j \tag{2}$$

And finally, the weights are updated:

$$\Delta W_{ij} = \eta \cdot y_i \cdot sign(e_j) \cdot |e_j|^{p-1} \tag{3}$$

where p is a parameter related to the energy function that is used to match the probability density function and the learning rule.

Then, lateral connections [7] are derived from the RGD and applied to MLHL. The resultant neural model takes the form of a network in which the independent factors of a data set may be identified, but in such a way that some type of global ordering is captured in the data set.

As a result, the final connectionist architecture is called CMLHL. The lateral connections effect acts after the feed-forward step, but prior to the feedback step.

The final architecture is therefore as follows: a feed-forward step (Eq. 1) followed by lateral activation passing:

$$y_i(t+1) = \left[y_i(t) + \tau(b - Ay)\right]^+ \tag{4}$$

Then there is a Feedback step (Eq. 2) followed by the weight change described in Eq. 3. Where:

– τ represents the "strength" of the lateral connections

– b is the bias parameter

– A is a symmetric matrix used to modify the response to the data based on the relation between the distances between the output neurons

– η is the learning rate

3.2.1 Lateral connections

Lateral connections have been derived from the RGD [7], which is a modification of the standard Gaussian distribution in which the variables are constrained to be non-negative, enabling the use of non-convex energy functions. The standard Gaussian distribution may be defined by:

$$p(y) = Z^{-1} \cdot e^{-\beta \cdot E(y)} \tag{5}$$

$$E(y) = \frac{1}{2} \cdot y^T \cdot Ay - b^T \cdot y \tag{6}$$

in which the quadratic energy function $E(y)$ is defined by the vector b and the symmetric matrix A. The parameter $\beta = 1/T$ is an inverse temperature. Lowering the temperature concentrates the distribution at the minimum of the energy function.

The cooperative distribution has been chosen in this study, as its modes are closely spaced along a non-linear continuous manifold. The sorts of energy function that can be used are those that block the directions in which the energy diverges to negative infinity. For this reason, matrix A has to fit the following property:

$$y^T \cdot Ay > 0 \quad \text{for all } y : y_i > 0, \quad i = 1 \ldots N \tag{7}$$

where N is the dimensionality of y.

The cooperative distribution in the case of N variables is defined by

$$A_{kp} = \delta_{kp} + \frac{1}{N} - \frac{4}{N} \cdot \cos\left(\frac{2\pi}{N}(k-p)\right) \text{ and} \tag{8}$$

$$b_k = 1 \tag{9}$$

where δ_{kp} is the Kronecker delta and k and p represent the output neuron identifiers.

Matrix A is used to modify the response to the data based on the relation between the distances between the outputs.

The projected gradient method is used, which consists of a gradient step followed by a rectification:

$$y_i(t+1) = \left[y_i(t) + \tau\left(b - Ay\right)\right] \tag{10}$$

where the rectification $[]^+$ is necessary to ensure that the y-values remain in the positive quadrant. If the step size τ is correctly chosen, this algorithm will probably be shown to converge to a stationary point of the energy function [*34*]. In practice, this stationary point is generally a *local minimum*.

The distribution mode can be approached by gradient descent on the derivative of the energy function with respect to y. This is expressed in Eq. 11:

$$\Delta y \propto -\frac{\partial E}{\partial y} = -\left(Ay - b\right) = b - Ay \tag{11}$$

The resultant network [*10*] can show the independent factors of a data set in a way that captures some type of global ordering in the data set and displays it with greater sparsity than other models.

Several versions of this model have successfully been applied to different real data sets, such

as data sets on banking, asteroid or algae classification [*6*], intrusion detection systems [*30*] and intelligent processing [*36*].

3.2.2 Fine-tuning

The CMLHL fine-tuning process is based on the effect of changing the τ parameter, which is the strength of the lateral connections between the output neurons. Experiments were conducted [*6*] using the bars data set proposed by [*46*] which adds noise in a graduated manner across the outputs. These experiments showed that altering the strength of the lateral connection parameter modulated the ability of the neural network to "gather" features together on the outputs. As predicted, a low τ value allows the neural model to code horizontal and vertical bars around a mode. An increase in the τ value means that the weak correlations between horizontal and vertical bars begin to have an impact on the learning. As the strength of the lateral connections becomes stronger, the bars are still trained around a mode but at the same time orientations start to separate. Subsequently, a separation emerges between the two different orientations, which is an interesting issue, because all the data inputs to the network consist of both horizontal and vertical bars. Increasing the τ value further forces the network to learn only one orientation of bars.

However, if the lateral connections are too strong, then the coding of the bars may be squashed into an area of the output space that is too small for all of the bars to be coded individually. The reason why one orientation of bars is suppressed is due to the pixel overlap between different orientations of bars. If the lateral excitation between the output neurons is strong enough, a single output neuron may be able to switch its preference from a horizontal bar to a vertical one. [*6*] considered orientation identification to be a precursor to the creation of horizontal/vertical concepts in animals inhabiting a mixed environment.

---

4 Case study: spectroscopic stained glass

The data under analysis are taken from an X-Ray spectroscopic analysis of a rose window in the west front of Burgos Cathedral (Spain), on the Way of St. James, the medieval pilgrimage to the tomb of St. James at Santiago de Compostela (the Cathedral was declared a World Heritage Site by UNESCO in 1984, and the Way of St. James, in 1993).

The aim of this research was to classify the spectroscopic data set and identify the sample's date and place of origin in order to facilitate the restoration of this and other historical stained glass windows. The same technique may also contribute to identifying different geographical areas in which the stained glass panels were produced, on the basis of their chemical composition and degree of conservation.

The main components of stained glass windows are composed of chemicals in different proportions such as silicon, sodium, potassium and calcium, which experts can use to determine the different origins of the glasses, and they may even go a step further by suggesting that the panes were made by different craftsmen. Typically, glass that is rich in sodium comes from coastal regions, whereas glass that is high in potassium comes from inland regions.

Spectroscopic techniques applied to minimal samples allow the characterization of historical materials without physical contact between the instrument and equipment, which ensures a non-destructive analysis and the integrity of the biopsy material under study. The technique is based on irradiating the sample material with a small laser. A spectrometer measures the light that spreads out and these data sets, among others, are used to determine the composition of the material.

The high-dimensional spectroscopic data set analysed in this study is composed of samples from 76 different panels from the aforementioned rose window. Its six colours are green, red, blue, yellow, pink and white. A morphological study revealed that the red-stained glass panels consisted of two layers: one of transparent glass and the other of coloured glass. The red-stained glass panel was therefore resampled as two separate samples. The data contained 450 data vectors obtained from 90 samples, each of which was analysed five times.

Each sample in the study was analysed five times, although not always under the same external conditions, which generated a significant lack of reproducibility due, for instance, to the different angles at which the beams penetrated the samples.

---

5 The unsupervised connectionist combined method

The high complexity and dimensionality of the spectroscopic data set investigated in this research led to the design of a three-step intelligent system (Fig. 1):

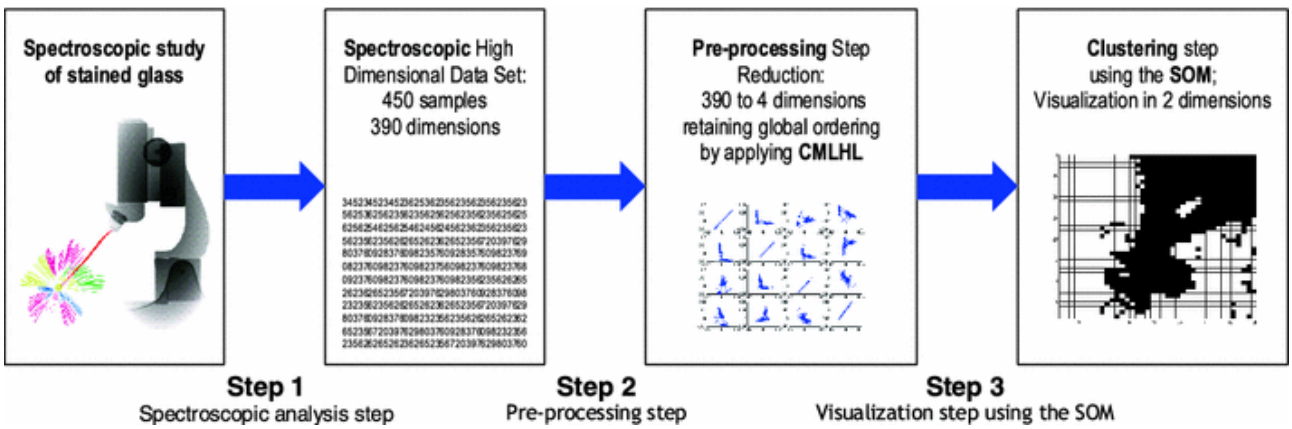

Fig. 1

Step 1 *Data acquisition by spectroscopic analysis and normalization step*. The original data (1,020 dimensions and 450 samples) was acquired after a spectroscopic analysis, as explained in the Sect. 4, which after normalization were reduced to 390 dimensions.

Step 2 *Pre-processing step*. A robust and fast pre-processing step was required to reduce the dimensionality and identify the most interesting features from the remaining 390 variables or dimensions. CMLHL is an EPP model applied in this study due to its capacity to capture some type of global ordering in the data set, based on the use of cooperative lateral connections. In this case, it performed a dimensionality reduction from 390 to 4 dimensions.

Step 3 *Visualization step*. An intelligent processing step, by which the data obtained in the previous step is analysed through a SOM. It provides a dimensionality reduction and visualization of the internal structure of the data set in two dimensions, where is easier to identify existing clusters and relations. This visualization method is applied in this research because each sample is analysed 5 times and not always under the same external conditions due, for example, to the different directions and angles of the beams that penetrate the samples. This implies an important lack of reproducibility. The SOM is a visualization method known to provide a low dimensional representation of a data set which captures local topographical relations based on the use of a neighbourhood function. Nearby neurons then quantize similar parts of the input space and similar inputs are quantized to the same or similar outputs. From the standpoint of analytical chemistry, this property is very important and, in combination with CMLHL, may allow us to solve or at least reduce the poor reproducibility of the data set under study based on differences in the external measurement conditions.

A graphical representation of the three-step unsupervised intelligent model proposed in the present study is shown in Fig. 1.

The novel unsupervised intelligent model was initially applied to a data set described by 1,020 dimensions and 450 samples, in order to provide a final visualization of the internal structure in two dimensions where clusters and relations are easier to identify and extract conclusions. To achieve this output, an initial normalization step was necessary from 1,020 dimensions to 390 dimensions. Then a fast and robust feature selection (pre-processing) step by CMLHL performed a dimensionality reduction from 390 dimensions to four dimensions. Both unsupervised neural models, the CMLHL and SOM, were applied due to their capabilities to capture local topographical relations, which together help to identify similar inputs that may serve to reduce the poor reproducibility effect of the chemical analysis in use and its negative impact.

6 Experiments and results

6.1 Pre-processing step

For comparison purposes, two models were applied to carry out Step 2 with the aim of performing a robust and fast dimensionality reduction from 390 to 4 dimensions, identifying the most interesting or relevant features. Fig. 2 shows the respective comparisons of the PCA and the CMLHL spectroscopic data projections over the first four eigenvectors/factors, in the form of a scatter plot matrix. It can be seen how the first model performs better than PCA (Figs. 2, 3). The vertical and horizontal axes forming these projections are combinations of the variables contained in the original datasets.

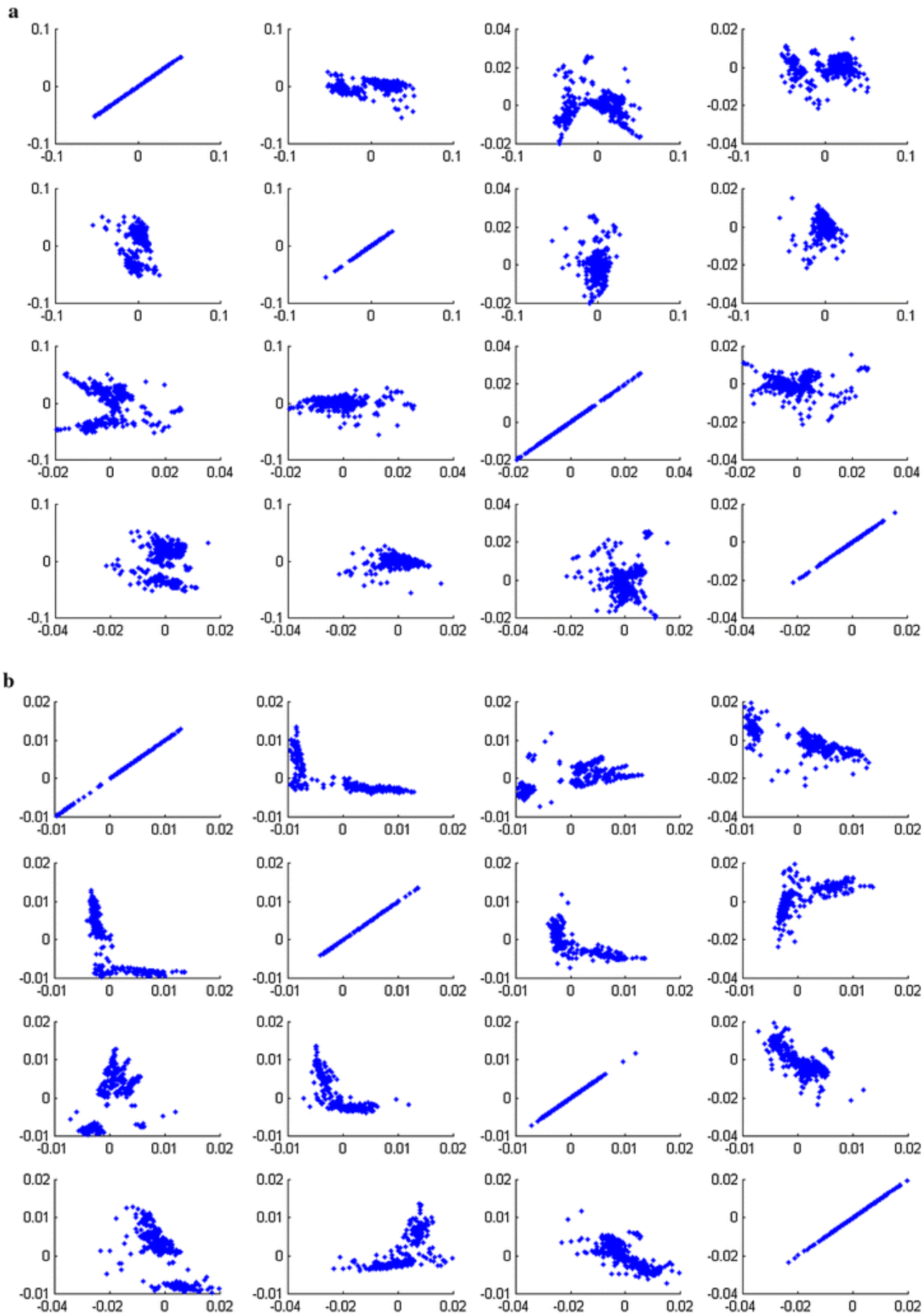

Fig. 2

Projections of the data set after a pre-processing step using CMLHL and PCA. **a** CMLHL projections on spectroscopic data—First four vector pairs. **b** PCA projections on spectroscopic data—First four vector pairs

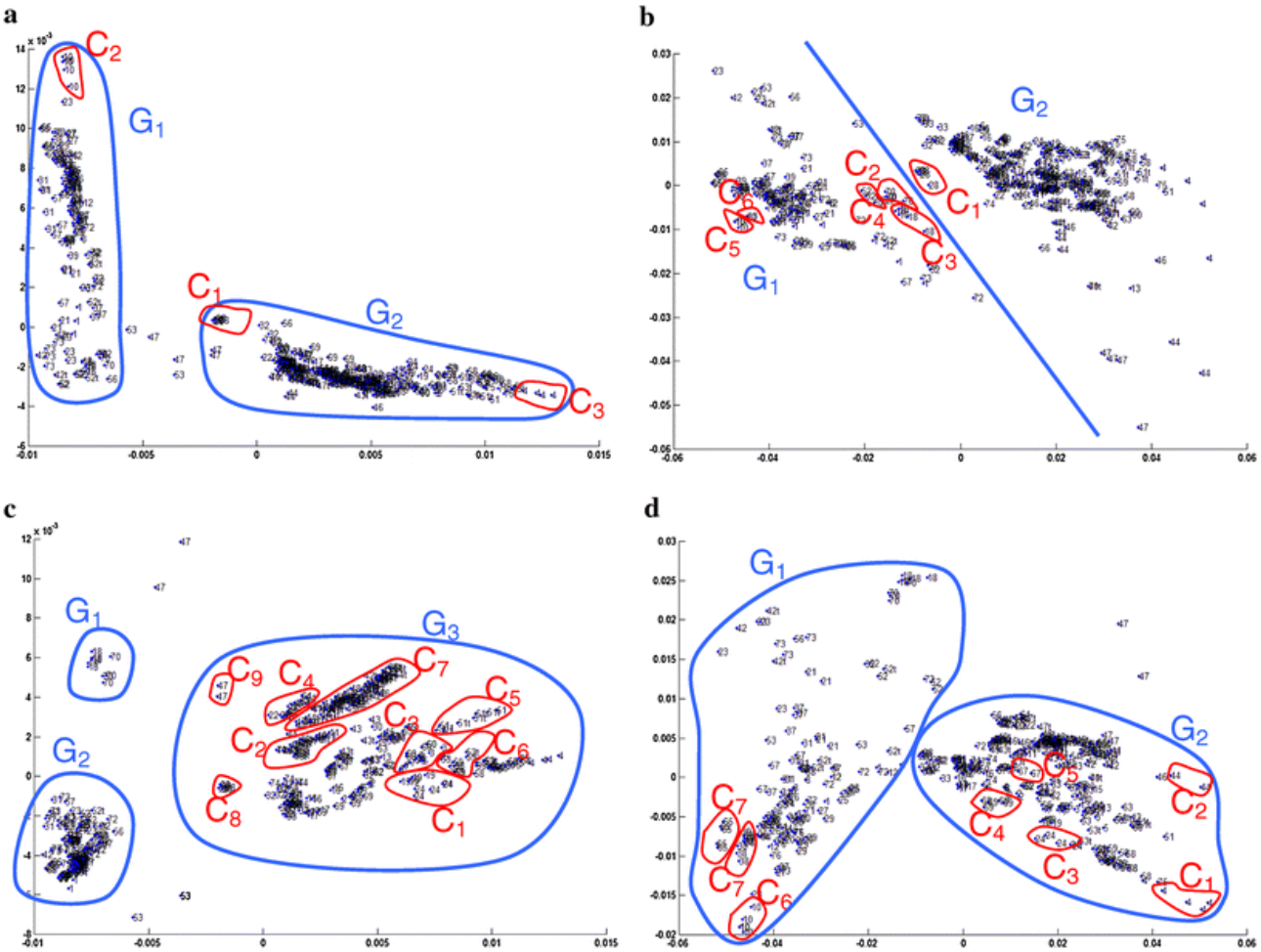

Fig. 3

A comparison of the first two PCA eigenvector pairs (Fig. 3b, c) against the first two CMLHL factor pairs (Fig. 3a, c). **a** CMLHL factor pair 1–2. **b** PCA eigenvector pair 1–2. **c** CMLHL factor pair 1–3. **d** PCA eigenvector pair 1–3

Figure 2a shows the application of a scatter plot matrix under the aforementioned connectionist model analysing the first four vectors pair-wise, projecting in each case the relation between two directions or dimensions. PCA (Fig. 2b) is applied, in this research for comparative purposes.

After the pre-processing step, some patterns can be already visualized [*19*]. CMLHL (Fig. 2a) clearly identifies more structure and greater separation between clusters than PCA (Fig. 2b). Figure 3 presents a more detailed comparison of the best projections, which are CMLHL factor 1-2, 1-3 pair and PCA eigenvector 1-2, 1-3 pair. It shows that both methods are able to identify some degree of internal structure based on the existence of several groups or clusters (two or three depending on the projections) and several sub-clusters.

In Fig. 3a, the projection obtained by CMLHL is more spread out with a larger separation between the two main groups G_1 and G_2, than that in the PCA projection (Fig. 3b). A general analysis leads to some interesting conclusions. For example, in Fig. 3a, there is a very strongly grouped central sub-cluster C_1, which contains class 28 class, sub-cluster C_3 to the right of group G_2 contains class 4 glass. Class 10 glass is found at the top of group G_1 in sub-

cluster C_2 .

The first eigenvector pair in Fig. 3b (PCA) shows two different groups G_1 and G_2, just as the CMLHL method does. In the centre, between these two groups, we can see two classes quite close, but separated by a straight line: sub-cluster C_1 relates to class 28; sub-cluster C_3 to class 18; and sub-clusters C_2 and C_4 to classes 70 and 52, respectively, all of which belong to group G_1. A clear structure based on the identification of two groups and sub-clusters is therefore identifiable. Unlike Fig. 3a, class 10 in the sub-cluster C_5 of group G_1 is not completely different from class 23 in sub-cluster C_6, which is spread throughout the sub-cluster from the top to the bottom of group G_1.

CMLHL factor pair 1-3 (Fig. 3c) is more defined than any other eigenvector/factor pair, and was able to show more structure as it identified three groups—G_1, G_2, and G_3—formed by many sub-clusters. For instance, group G_1 is composed of class 18 and 70, both of which are very compact and linearly separable from each other. Group G_3 is formed by many sub-clusters such as for example C_8 (class 28) and class 47 is related to sub-cluster C_9 . So, it can be seen that Group G_3 presents quite an interesting structure made up of several sub-clusters. Some examples of this structure of group G_3 are the nine sub-clusters C_1–C_9 identified in Fig. 3c. Table 1 shows some of the classes in each of these sub-clusters.

Table 1

Classes belonging to six of the sub-clusters found in group G_3 (Fig. 3c)

Cluster	Classes
C_1	19, 24
C_2	11, 40, 70
C_3	3, 7, 60
C_4	22, 64
C_5	51, 51t
C_6	53, 59

PCA eigenvector pair 1-3 (Fig. 3d) identified some clustering in group G_2, but upon investigation these sub-clusters were not very well defined, and they contained some sub-clusters but mainly a mix of classes with a high content of potassium. In the other cluster, group G_1, the clustering is even less defined, groups G_1 and G_2 being less defined than in the case of CMLHL (Fig. 3c).

In the case of Figs. 2 and 3 and in general for PCA and CMLHL projection models, the vertical and horizontal axes forming these projections are combinations of the variables contained in the original data sets. These figures therefore show the projections of different vector pairs. In the case of this specific data set they are based on the chemical composition of the glasses samples such as the potassium and sodium content, and so on.

6.2 Visualization step

In order to obtain a final and more detailed description of the internal structure of this high-dimensional data set, a dimensionality reduction step has been applied, going from four dimensions to visualization in two dimensions, by applying a SOM. The results obtained are presented in Fig. 4. Finally several groups can be clearly identified. An analysis from a chemical composition point of view revealed that the group G_1, formed by the sub-clusters C_1–C_5 (Fig. 4), comprises samples with high sodium contents and those with high potassium contents are found in group G_2, formed by the sub-clusters C_6–C_{11} (Fig. 4).

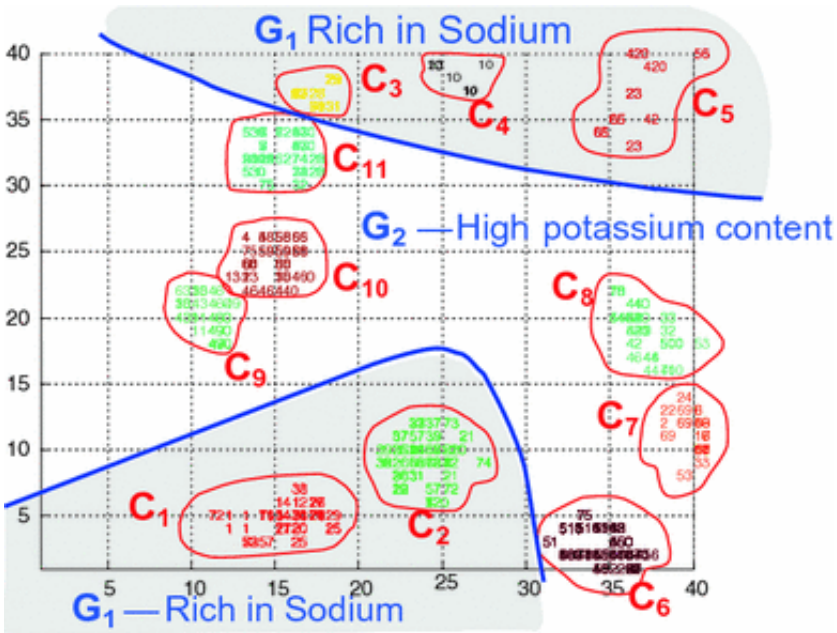

Fig. 4

A SOM visualization using the proposed pre-processing step, using an initial CMLHL feature selection step

Figure 4 shows the result obtained by the SOM in a 2D input space. This figure depicts the grid of the model (40 × 40 neurons) embedded over the data set, providing an at-a-glance view of the adaptation of the map to the data set based on the chemical composition of the glasses samples, which is mainly potassium and sodium content in this case of study.

It can be seen (Fig. 4) how the novel model presented in this study is able to classify the different samples in well-defined groups and sub-clusters, based mainly on chemical composition, revealing a classification of groups that are rich in either sodium (G_1) or

potassium (G_2). A detailed classification giving information about how each group is formed is shown in Table 2.

Table 2

shows the crystal samples identified within each group and the number of times that each sample has been identified out of five. For the sake of simplicity, only samples identified four or five times out of five are shown

Sub-cluster	Sample id (number of times that each sample is identified out of five)
G_1–Sub-clusters rich in sodium	
C_1	1 (5), 20 (4), 25 (5), 34 (5), 76 (5)
C_2	8 (5), 12 (4), 21 (5), 26 (5), 35 (5), 36 (5), 37 (5), 38 (4), 39 (5), 57 (4), 68 (5), 72 (4)
C_3	29 (4)
C_4	10 (5)
C_5	65 (5)
G_2–Sub-cluster rich in potassium	
C_6	22 (5), 45 (4), 48 (5), 51 (5), 55 (5), 64 (5), 45t (5), 47t (4), 48t (5), 51t (4), 55t (5), 56t (5)
C_7	6 (5), 16 (5), 17 (5), 61 (5), 69 (5)
C_8	18 (5), 33 (4), 41 (4), 42 (4), 44 (4), 54 (5), 70 (4), 44t (4), 50t (5), 54t (5)
C_9	11 (5), 30 (4), 49 (4), 46t (4), 49t (4)
C_{10}	4 (4), 58 (5), 59 (5), 60 (5), 66 (5)
C_{11}	3 (4), 5 (4), 7 (5), 19 (5), 24 (4), 28 (4), 62 (4), 67 (4), 74 (4), 43t (4), 53t (5)

It can be seen that 68 out of 76 samples were identified four or five times from among the five separated analyses.

It should be noted that the difficulty of reproducing the experimental conditions of an X-ray spectroscopic analysis is due to existing anomalous samples. In this case, only eight samples were identified as having this problem. The main part of the reproducibility information lies in a part of the signal on the continuous spectrum; thus, it may only be obtained from the continuous spectrum. One of the main advantages of using machine learning techniques such

as ANN is that they have the capacity to generalize. Moreover, the use of a topological preserving mapping model such as the SOM and CMLHL helps to reduce the problem of poor reproducibility.

The final result is due to the different proportions of the main components such as silicon, sodium, potassium and calcium, which are the different glass components, and more importantly, the different melt materials. The main difference between groups made up of original glasses is in this case their sodium (G_1) and potassium (G_2) content (Fig. 4). This fact shows that the raw materials of the different glasses have different origins, and we could even go further by saying that they were made by different craftsmen. In general, glass that is rich in sodium, (group G_1 formed by the sub-clusters C_1 to C_5, from Fig. 4) comes from coastal regions where seaweed ash was used for its production, whereas glass which is high in potassium content (group G_2 formed by the sub-clusters C_6 to C_{11}, from Fig. 4) comes from inland regions where wood ash was used.

There is no any relation between the colour factor and the groups; in other words, no glass with the same colour appears indistinctly in the different groups. All the glasses which belong to the same group come from the same common matrix and the differences arise from the addition of transition metals in small concentrations, which were used as colouring materials.

The materials which show a high level of corrosion belong to the category in which the potassium stands out as another constituent. On the other hand, the samples which belong to the group that reveals greater sodium composition were kept in a better state of conservation.

Among the glasses with the same concentration of alkaline components are those with lithium, followed by others with sodium and finally less stable ones with potassium. So, there is a high correlation between their chemical composition and degree of conservation.

It is important to bear in mind that the groups are influenced by the poor reproducibility of the chemical analysis in use. In principle, there would be five similar results if the technique were reproducible that would appear alongside each other. Thus, the samples that appear to be separated and identified with the same number represent outliers. One way of improving on these results, which we currently consider an area of future work, would be to filter the data, which implies the elimination of outliers or isolated samples. In our case it will be those that only appear once or twice, or even three times in a sub-cluster. The elimination of these anomalous samples and the application of the proposed method would, in our opinion, define these groups better, as only the chemical information that is correctly obtained would be present in the reproducibility information relating to the analytical technique.

---

7 Comparison with some other methods

The proposed three-step unsupervised connectionist model provides more information in terms of visualization than the results obtained by PCA or CMLHL (Figs. 2, 3) alone as previously demonstrated in Sect. 6. This model produces a better analysis as it identifies a greater number of well-defined groups and subclusters.

Figure 5a shows the SOM error, having used CMLHL as a preprocessing method, as the number of iterations increases. It may be clearly seen in Fig. 5a that there is a marked tendency for the error to decrease, and it is therefore likely that it will achieve convergence.

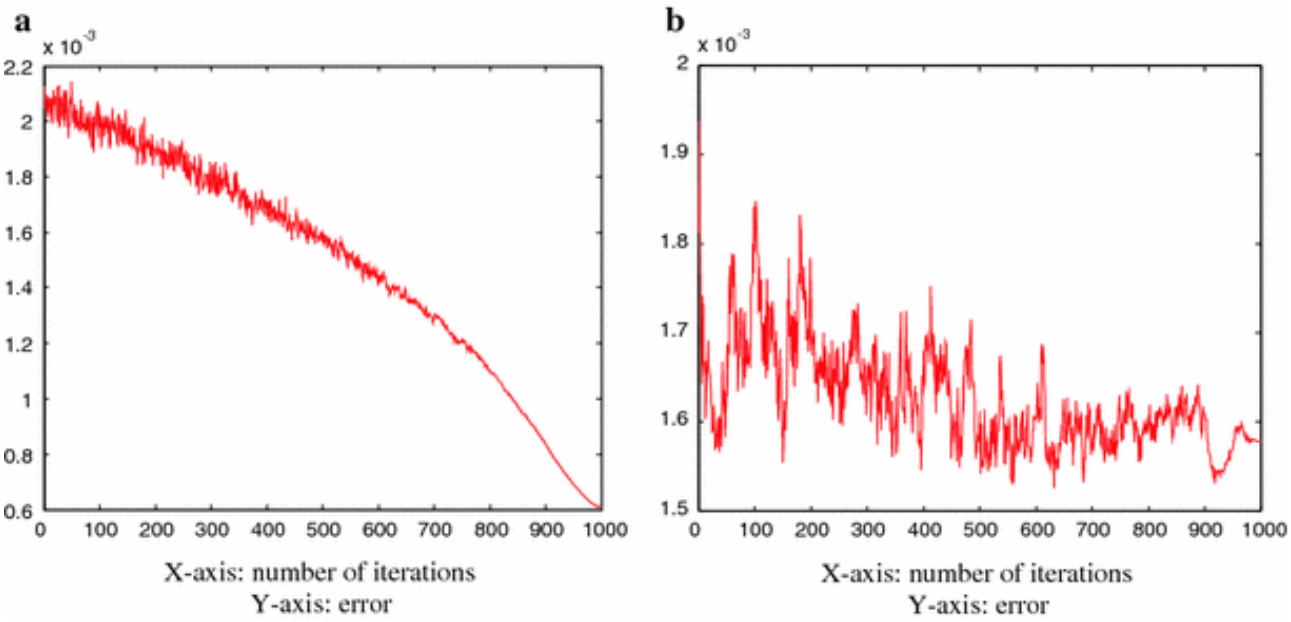

Fig. 5

Error behaviour of the models as the number of iterations increases. **a** Representation of the error versus the number of iterations in the case of the CMLHL/SOM model proposed in this study. **b** Representation of the error versus the number of iterations using a classical SOM

The same kind of analysis has been applied to this data set, but this time only applying the SOM without any pre-processing step. In this case the behaviour of the error (Fig. 5b), which differs greatly from the earlier case, shows how convergence is much more erratic without any pre-processing step.

7.1 Conclusion and future work

In this study, we have presented a novel three-step unsupervised neural method using the CMLHL as a feature selection or pre-processing step, followed by the application of the SOM to analyse complex high-dimensional spectroscopic data characterized by their lack of reproducibility. It has been shown how CMLHL performs much better than a classical model such as PCA as a pre-processing method. CMLHL has also been applied due to the action of its lateral connections in order to preserve a certain degree of topology as the SOM does. The proposed model has identified an interesting visualization of the different stained glass panels

solving the difficulty caused by the differences in external test conditions.

The performance of the combined visualization neural model is shown to be better than the performance of PCA or CMLHL alone, when analysing the internal structure of the data set.

It has also shown its fast convergence ability in comparison with a standard SOM.

Future work will be based on the refinement of the initial data set based on the identification of outliers using, for instance, re-sampling techniques [*31*]. This combined model will also be applied to other data sets that are of interest. A further extension would be the application of different versions of the SOM, such as the ViSOM [*32*, *33*] and WeVoS-SOM [*35*], in order to obtain better visualization results.

Acknowledgments

This research was supported by projects TIN2010-21272-C02-01 from the Spanish Ministry of Science and Innovation and BU006A08 of the JCyL. The authors would also like to thank the manufacturer of components for vehicle interiors, Grupo Antolin Ingeniería, S.A. which provided support through MAGNO 2008 – 1028 – CENIT funded by the Spanish Ministry of Science and Innovation.

References

1. Ahmad A, Dey L (2005) A feature selection technique for classificatory analysis. Pattern Recogn Lett 26(1):43–56
CrossRef

2. Kohonen T (1988) Self-organisation and associative memory, vol 8, Springer series in information sciences. Springer-Verlag, New York

3. Erwin E, Obermayer K, Schulten K (1992) Self-organizing maps: ordering convergence properties and energy functions. Biol Cybern 67:47–55
MATH CrossRef

4. Wiskott L, Sejnowski TJ (1998) Constrained optimization for neural map formation: a unifying framework for weight growth and normalization. Neural Comput 10(3):671–716
CrossRef

5. Svensen M (1999). The generative topographic mapping PhD thesis. Aston University, UK

6. Corchado E, MacDonald D, Fyfe C, (2004). Maximum and minimum likelihood Hebbian

learning for exploratory projection pursuit. Data mining and knowledge discovery. Kluwer Academic Publishing 8(3):203–225

7. Seung HS, Socci ND, Lee D (1998) The rectified Gaussian distribution. Advances in neural information processing systems 10:350

8. Laaksonen J, Koskela M, Laakso S, Oja E (2001) Self-organising maps as a relevance feedback technique in content-based image retrieval. Pattern Anal Appl 4(2–3):140–152
MathSciNet MATH

9. Lagus K, Kaski S, Kohonen T (2004) Mining massive document collections by the WEBSOM method. Inf Sci 163(1–3):135–156
CrossRef

10. Corchado E, Fyfe C (2003). Connectionist techniques for the identification and suppression of interfering underlying factors. International journal of pattern recognition and artificial. Intelligence. 17(8):1447–1466

11. Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philos Mag 2:559–572

12. Hotelling H (1993) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–444
CrossRef

13. Fyfe C, MacDonald D (2002) Epsilon-insensitive Hebbian learning. Neurocomputing 47(1–4):35–57
MATH

14. Ahmadi A, Omatu S, Kosaka T (2003) A PCA based method for improving the reliability of bank note classifier machines. In: Loncaric S, Neri A, Babic H (eds), ISPA 2004 Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis (IEEE Cat. No. 03EX651), vol 1. Univ. of Zagreb, Zagreb, Croatia, pp 494–499. doi:10.1109/ISPA.2003.1296947

15. Hyvärinen A (1997). New approximations of differential entropy for independent component analysis and projection pursuit. NIPS 1997

16. Diaconis P, Freedman D (1984) Asymptotics of graphical projections. Ann Stat 12(3):793–815
MathSciNet MATH CrossRef

17. Sanger D (1989) A technique for assigning responsibilities to hidden units in

connectionist networks contribution analysis. Conn Sci 1(2):115–138
CrossRef

18. Demtröder W (2008) Laser spectroscopy: experimental techniques, 4th edn. Springer, Berlin

19. MacDonald D, Corchado E, Fyfe C et al. (2003). Maximum-likelihood competitive learning for the analysis of spectroscopic data. 2nd International Workshop on Practical Applications of Agents and Multiagent Systems–IWPAMS 2003

20. Yang HC, Lee CH (2004) A text mining approach on automatic generation of web directories and hierarchies. Expert Syst Appl 27(4):645–663
MathSciNet CrossRef

21. Yang HC, Lee CH (2004) Mining text documents for thematic hierarchies using self-organizing maps. Comput Rev 45(2):117–118
MathSciNet

22. Yang HC, Lee CH (2005) A text mining approach for automatic construction of hypertexts. Expert Syst Appl 29(4):723–734
CrossRef

23. Kohonen T (2000) Data mining by the self-organising map method. In: Bouchon-Meunier B, Yager RR, Zadeh LA (eds.) Uncertainty in intelligent and information systems. Advances in fuzzy systems—applications and theory, vol 20. World Scientific, Singapore, pp 3–22

24. Abonyi J, Nemeth S, Vincze C, Arva P (2003) Process analysis and product quality estimation by self-organizing maps with an application to polyethylene production. Comput Ind 52(3):221–234
CrossRef

25. Lessmann B, Degenhard A, Kessar P, Pointon L, Khazen M, Leach M O, Nattkemper T W (2005). SOM-based wavelet filtering for the exploration of medical images. In: Artificial neural networks: biological inspirations–ICANN 2005, Pt. 1, Proceedings, Lecture Notes in Computer Science, pp 671–676

26. Krell G, Rebmann R, Seiffert U, Michaelis B (2003). Improving still image coding by an SOM-controlled associative memory. In: Sanfeliu A, Ruiz-Shulcloper J (eds.) Progress in pattern recognition, speech and image analysis. 8th Iberoamerican Congress on Pattern Recognition, CIARP 2003. Proceedings Lecture Notes in Computer Science. Springer-Verlag, Berlin, pp 571–579

27. Lin S, Si J (1998) Weight-value convergence of the SOM algorithm for discrete input. Neural Comput 10(4):807–814
CrossRef

28. Corchado JM, Aiken J, Corchado E, Fernández F (2005) Evaluating the air-sea interactions and fluxes using an instance-based reasoning system. AI Communication 18(4):247–256
MATH

29. Herrero A, Corchado E, Pellicer MA, Abraham A (2009) MOVIH-IDS: a mobile-visualization hybrid intrusion detection system. Neurocomputing 72(13–15):2775–2784
CrossRef

30. Herrero A, corchado E, Gastaldo P, Zunino R (2009) Neural projection techniques for the visual inspection of network traffic. Neurocomputing 72(16–18):3649–3658
CrossRef

31. Bogdan G, Baruque B, Corchado E (2006) Outlier resistant PCA ensembles. In: Knowledge-based intelligent information and engineering systems, 10th international conference, KES 2006, Bournemouth, UK. KES. LNAI, vol. 3. Springer, Heidelberg, pp 432–440

32. Yin H (2002) Data Visualisation and Manifold Mapping Using the Visom. Neural Networks 15:1005–1016
CrossRef

33. Baruque B, Corchado E (2007) Fusion of visualization induced SOM. Innovations in hybrid intelligent systems series: advances in soft computing, vol 44. Springer, Berlin

34. Bertsekas DP (1995) Nonlinear programming. Athena Scientific, Belmont
MATH

35. Baruque B, Corchado E (2010). A weighted voting summarization of SOM ensembles. Data mining and knowledge discovery. Springer. 21(3):398–426. doi:10.1007/s10618-009-0160-3

36. Herrero A, Corchado E, Sáiz L, Abraham A (2010) DIPKIP: a connectionist knowledge management system to identify knowledge deficits in practical cases. Comput Intell 26(1):26–56
CrossRef

37. Yan W, Chen CH, Khoo LP (2005) A web-enabled product definition and

customization system for product conceptualization. Expert Syst 22(5):279–293
CrossRef

38. Liu H, Liu L, Zhang H (2009). Boosting feature selection using information metric for classification. In: Neurocomputing. vol 73(1–3). Elsevier Science, Amsterdam

39. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics, vol 23(19). Bioinformatics Oxford University Press, Oxford, pp 2507–2517

40. Vinaya V, Bulsara N, Gadgil CJ, Gadgil M (2009) Comparison of feature selection and classification combinations for cancer classification using microarray data. Int J Bioinform Res Appl 5(4):417–431
CrossRef

41. Nemati S, Basiri ME, Ghasem-Aghaee N, Aghdam MH (2009) A novel ACO-GA hybrid algorithm for feature selection in protein function prediction. Expert Syst Appl Int J 36(10):12086–12094
CrossRef

42. Hua J, Tembe WD, Dougherty ER (2009) Performance of feature-selection methods in the classification of high-dimension data. Pattern Recogn 42(3):409–424
MATH CrossRef

43. Gunter S, Bunke H (2004). An evaluation of ensemble methods in handwritten word recognition based on feature selection. Pattern Recogn. ICPR 2004

44. Gunter S, Bunke H (2004) Handwritten word recognition using classifier ensembles generated from multiple prototypes. Int J Pattern Recogn Artif Intell 18(5):388–392

45. Sun NQ, Li Y (2009) Intrusion detection based on back-propagation neural network and feature selection mechanism. FGIT 2009. LNCS 5899:151–159

46. Földiák P (1991) Models of sensory coding, PhD dissertation, University of Cambridge (reprinted as Technical Report No. CUED/F-INFENG/TR 91, Department of Engineering, University of Cambridge, 1992)

47. Khuwaja GA (2005) Merging face and finger images for human identification. Pattern Anal Appl 8:188–198
MathSciNet CrossRef

48. Hurtado L F, Griol D, Segarra E, Sanchís E (2006) A stochastic approach for dialog management based on neural networks. In: Proceedings of the 9th international conference on spoken language processing interspeech, Pittsburgh, pp 49–52

49. Chow TWS, Rahman MKM, Wu S (2006) Content-based image retrieval by using tree-structured features and multi-layer self-organizing map. Pattern Anal Appl 9:1–20
MathSciNet CrossRef