DOI 10.5162/IMCS2012/P1.9.16

Nonlinear Algorithms to Reduce the Dimension of Databases without Loss of Information

José S. Torrecilla, Gemma Matute, Carlos Calvo, Claudia Ceña, Francisco Rodríguez
Department of Chemical Engineering. Complutense University of Madrid, 28040 Madrid, Spain.
(Telephone: +34 91 394 42 44; Fax: +34 91 394 42 43; email: jstorre@quim.ucm.es)

Abstract

In this work, self-organizing maps have been used in the reduction of the dimensionality of the database, extracting the essential information of the database, facilitating its handling and reducing the time needed by the sensor to give the measurement. This tool has been applied to a database composed of 220 1H NMR and 31P NMR spectra of 13 using edible vegetal oils (hazelnut, sunflower, corn, soybean, sesame, walnut, rapeseed, almond, palm, groundnut, safflower, coconut, and extra virgin olive oils). With this tool, the dimension of the databases decreases from 11 x 192 to 2 x 192. The loss of information was checked by comparing the statistical results shown here with others that can be found in literature. Here, using a low dimension database and without any other physicochemical feature, the statistical results have been slightly improved (the misclassification percentage decreases from 3 to less than 2.8%).

Key words: Dimensionality reduction; Self-organizing map; Edible oils.

Introduction

To understand what a sensor is, our five senses give excellent examples. Everybody has five groups of sensors, and in each one there are connections between the biological detectors and the brain, for instance, in the case of the olfactory sense, the information taken in the olfactory membrane is transmitted through a nerve cell to be treated in the brain. The brain perceives, encodes, and recognizes the received information through the senses. In artificial sensors, similar tasks are carried out. And therefore, to improve the sensor performance, the development involves the progress of every part of the sensor. In all cases, the high dimension of the database produced by the sensor makes their handling difficult inside the device. Undoubtedly, the control or even the measurement of some physicochemical properties from complex chemical processes usually produces databases with high dimension. In addition, it is known that with the increase in volume, complexity and dimensionality of the databases, their analysis becomes progressively more cumbersome. Because of this, it is necessary to reduce the volume and dimensionality of the databases without the loss of essential information.

Traditional methods used to reduce the dimensionality of databases are linear algorithms such as principal component analysis (PCA), independent component analysis (ICA) or classical multidimensional scaling (MDS) [1,2]. The common weakness of these tools is that they are suitable only under linear constraints. Taking into account that in real systems these limitations are not habitually achieved, nonlinear algorithms have been tried here.

The aim of this work is the application of a self-organizing map as a nonlinear tool to reduce the dimensionality of databases used to classify 13 edible vegetable oils (hazelnut, sunflower, corn, soybean, sesame, walnut, rapeseed, almond, palm, groundnut, safflower, coconut, and extra virgin olive oils) without loss of the essential information.

Materials and Methods

The database used to design and optimize the self-organizing map model consists of values of the acidity, iodine value, ratio of 1,2-diglycerides to the total diglycerides and the concentrations of total sterols, total diglycerides, 1,2- diglycerides, 1,3-diglycerides, saturated fatty (SFA), oleic, linolenic, and linoleic acids determined by analysis of the respective 1H NMR and 31P NMR spectra [3]. These properties were calculated in 220 samples corresponding to 13 types of vegetables oils (hazelnut, sunflower, corn, soybean, sesame, walnut, rapeseed, almond,

palm, groundnut, safflower, coconut, and extra virgin olive oils). These samples have respectively been distributed in 192 and 28 samples in the learning and verification samples. Therefore the initial dimension is 11 x 220.

In order to guarantee the reliability of the classifications carried out by this model, the applicability domain has been evaluated selecting the compounds with cross-validated standardized residuals greater than three standard deviation values. The widest dispersion is presented in the SFA data (<3.6 standard deviation), where the coconut and almond oils present the highest and the lowest ranges of values, respectively. As these oils form two of the 13 types of oils, these samples are not considered as outliers. The mean repeatability and reproducibility values of the databases used are less than 2 and 2.5%, respectively [3].

Self-Organizing Maps (SOMs) or Kohonen neural networks are abstract mathematical models of the mapping between the sensory nerve and the cerebral cortex. It is one of the most interesting topics in the competitive and non-supervised neural network field [4]. SOM models can learn to detect irregularities, correlations in their inputs, adapt their future responses to that input accordingly and classify input data depending on how they are grouped in the input space. These types of maps are able to recognize groups with similar characteristics [4,5].

The architecture of SOM models is shown in Figure 1. Every circle and arrow represent a neuron and weight, respectively; that is, there are as many weights as arrows and the number of neurons is equal to the product of the width and length of the competitive layer. In this layer, each neuron has as many weights as the input descriptors (concentration of SFA and oleic acid). Every neuron is represented by a vector of weights. The self-organizing map used in this work was design using Matlab version 7.01.24704 (R14) [5].

Type Style and Size

Text should be single-spaced in Arial typeface 10 pt.. We recommend using the defined style “IMCS_Bodytext”.

Illustrations, Graphs, and Photographs

Please arrange figures in a way that important information can be caught at a glance. Illustrations, graphs, and photographs should fit preferably in one column (see Fig. 1), but for larger graphs, also two-column figures are possible (see Fig. 2). But please keep the text in the two-column format! The resolution for photographs and graphics should not exceed 300 dpi.

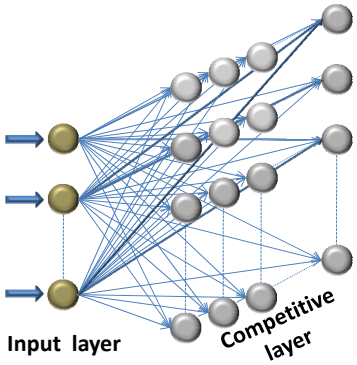

Fig. 1. Schematic diagram of self-organizing map model.

Results and discussions

The main objective is to classify each of the 13 edible oil types using the least number of independent variables. That is, the initial dimension (11 x 192) has to be reduced, Figure 2. Torrecilla et al. reduced the initial dimension to three independent variables (3 x 192) by PCA techniques and some important characteristics of the studied oils [6]. As a continuation of this research line, another self organizing map has been optimized and used to reduce even more the number of independent variables.

Fig. 2. General scheme of the work presented.

A nonlinear mapping method was used to classify the aforementioned 13 types of oils. The topology of this SOM consists of 11 input neurons (values of the acidity, iodine value, ratio of 1,2-diglycerides to the total diglycerides and the concentrations of total sterols, total diglycerides, 1,2-diglycerides, 1,3-diglycerides, SFA, oleic, linolenic and linoleic acids).

Self-Organizing Map optimization process: The output neurons were arranged in three different topological grids viz. grid, hexagonal and random topologies. In addition, two different methods to calculate the distances were used, viz. Euclidean and Manhattan distances [5]. Both topologies and distance were combined and the best pair was selected. The combination with the least number of misclassifications was selected. In this case, hexagonal topology and Manhattan distance were selected. Once the topology and distance of the SOM were selected, the dimension of the network was optimized. Networks of sizes

ranging from 18 × 18 to 26 × 26 were tried [6]. Taking into account the number of misclassifications, the output map dimension was 19 × 19. With the selected topology, distance and the optimized network dimension, the parameters of the SOM model were optimized by a Central Composite Design 25 + star experimental design, where the variables analyzed were ordering phase learning rate (from 0.1 to 1), ordering phase steps (from 500 to 1500), tuning phase learning rate (from 0.01 to 0.03), neighborhood distance (from 0.5 to 1.5) and the number of epochs in the learning process (from 10000 to 30000 epochs) [5]. The response of the experimental design was the number of incorrect classifications of the oil samples.

Throughout the learning process, the competitive neurons have been adequately distributed in whole three-dimensional space, and therefore, every data set would be classified by one or a group of neurons. The only misclassification consisted of a hazelnut sample which was classified as EVOO. This mistake is based on their similar chemical composition. In order to reach the least number of misclassifications, the optimum parameter values have been fixed at 0.1, 1500, 0.01, 0.5, and 30000 to ordering phase learning rate, ordering phase steps, tuning phase learning rate, neighborhood distance and the number of epochs necessary in the learning process, respectively [5].

Application of Self-Organizing Map: Once the parameters of SOM had been optimized, the results were analyzed. The input can be grouped in the output maps in two different regions characterized by two of the three most important variables shown in literature viz. oleic acid and SFA concentration [6]. With these two independent variables, the same calculations and internal validation process as in the mentioned work were carried out.

In this case, using the databases with the lowest dimensionality, the learning and verification samples have the same format. These have two rows as variables necessary to characterize the process (concentrations of SFA and oleic acid) and the same number of columns as the number of vectors to describe the system to be studied. Whole database has been distributed randomly into learning (80%) and verification (20%) samples. Using the low dimension database the misclassification percentage is less than 2.8 % (3% in the case of [6]). The statistical results achieved were slightly better than in the aforementioned work. Taking into account that no characteristic physicochemical information of the edible oil has been required, SOM is a suitable tool to reduce the dimensionality of the databases without appreciable loss of information. Therefore, in the light of these results, this type of map is able to classify nearly all input data used in the validation sample.

Conclusions

In this work, a mathematical approach based on self-organizing map networks models have been designed to reduce the dimensionality of databases, extracting the essential information of the database, facilitating its handling and reducing the time needed by the sensor to give the measurement. These points have been developed using only the concentration of two of the chemicals present in most vegetable oils (oleic acid and saturated fatty acids). With this tool, the dimension of the databases decreases from 11 x 192 to 2 x 192. The loss of information was checked by comparing the statistical results shown here with others that can be found in literature [6]. Here, using a low dimension database and without any other physicochemical feature, the statistical results have been slightly improved. Using the low dimension database the misclassification percentage has decreased from 3 to less than 2.8 % [6].

Acknowledgements

The research leading to these results has achieved funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. HEALTH-F4-2011-258868.

References

[1] Cox, T; Cox, M. Multidimensional Scaling. 2nd ed. London: Chapman and Hall/CRC; 2000.

[2] Stone, J. V. Independent Component Analysis: A Tutorial Introduction. Cambridge: The MIT Press; 2004.

[3] Vigli, G.; Philippidis, A.; Spyros, A.; Dais, P. Classification of edible oils by employing 31P and 1H NMR spectroscopy in combination with multivariate statistical analysis. A proposal for the detection of seed oil adulteration in virgin olive oils. J. Agric. Food Chem. 2003, 51, 5715-5722.

[4] Kohonen, T. Self-Organization and associative Memory, 2nd ed.; Berlin: Springer-Verlag, 1987.

[5] Demuth, H.; Beale, M.; Hagan, M. MATLAB User's Guide, V 4.0; Neural Network Toolbox, MathWorks Inc., Mass., USA, 2005.

[6] Torrecilla, J. S.; Rojo, E.; Oliet, M.; Domínguez, J.C.; Rodríguez, F. Self-Organizing Maps and Learning Vector Quantization Networks As Tools to Identify Vegetable Oils. J. Agric. Food Chem., 2009, 57, 2763-2769.