Non-supervised neural categorisation of near infrared spectra

Paloma Cáceres-Alonso
Bran+Luebbe, Madrid, Spain.

Roberto Rodríguez-Galán and Alvaro García-Tejedor
Computer Science Dept, Universidad Carlos III, Madrid, Spain.

Introduction

Raw materials identification by means of near infrared (NIR) spectroscopy can be seen as a generic pattern recognition problem. Thus, it can be approached using standard supervised pattern recognition techniques, either non-parametric (KNN, PRIMA) or parametric (LDA, SIMCA).[1] Nonetheless, these methods classify samples of unknown origin on the basis of a group of samples with an *a priori* known classification.

Over the last years a promising computational technique called artificial neural networks (ANN) has emerged in the world of near infrared spectroscopy opening up new possibilities mainly for non-linear data handling.[2,3] They do not rely on the existence of previous human experience, but extract relationships (both linear and non-linear) between data sets presented during the training process. Besides, neural networks have shown their ability to deal with changing information through generalisation and self-learning capabilities.

Most of ANN and NIR publications are devoted to the solution of the quantitative problem.[2] Only in some very recent papers has the use of this technique for qualitative analysis been found.[4–6] In both cases, multi-layer perceptrons (MLP), usually trained with the generalised Delta rule and the back-propagation algorithm, is the basic architecture. The drawback for back-propagation networks is that supervised training is required. Unsupervised pattern recognition methods appear as an alternative solution for substance identification (qualitative analysis). Self organising maps[7] (a kind of artificial neural networks with unsupervised training algorithm) have been revealed as an efficient and widely used categorisation method[8] due to its ability to build non-linear mappings between high-dimensional variable spaces and low-dimensional class spaces without external guidelines.

In this paper we present a non-supervised neural system that performs raw material identification using a pattern recognition technique based on Kohonen's self organising maps. The result is a tool that groups input substances depending on the resemblance and regularities of their spectra. Experiments to obtain the best combination of parameters have been carried out. The resulting classifier is tested with fresh data to evaluate the goodness of the model. Conclusions on these experiments are also discussed.

Experimental basis

NIR reflectance spectra have been recorded with a Fourier-transform spectrophotometer, Infraprover I (Bran+Luebbe GmbH), from 4500 to 10,000 cm^{-1}, in 25 cm^{-1} steps with a two metre

Table 1. List of substances and codes. C/T number of spectra used for calibration/test.

N	Code	Substance name	C	T
1	001	Barbituric acid	4	3
2	003	EDTA	2	2
3	004	D (+) Maltose	4	2
4	005	Sorbitol	4	3
5	006	Galactose	2	2
6	012	Vanillin	4	3
7	013	Sodium bisulphite	2	2
8	021	Sulfathiazole	4	4
9	023	Ethyl vanillin	4	4
10	028	Proprietary product	4	4
11	033	Proprietary product	4	1
12	043	Salicylic Ac. acetate	4	4
13	064	Metronidazole	2	1
14	068	Dry Al hydroxide gel	2	2
15	069	Wet Mg hydroxide gel	2	2

fiber optic bundle. The raw data obtained are divided by a standard ceramic reference mounted at the end of the fiber to be able to obtain reliable measurement, over the time. Every spectrum is an average of five readings. A group of 15 different solid materials have been chosen among the standard demo library used by B+L. The only condition when sampling the data was to obtain a broad representation of organic and inorganic compounds with low and high reflectance values. In general measurements of different lots are very reproducible, trying to avoid an increase in complexity. Some of the samples are proprietary and cannot be identified. The name of each spectrum has been coded: three digits represent the type of substance. Our training set have 48 spectra and the test set 39, with 459 spectral data per spectrum. Table 1 summarises the substance number and code and the number of C—Calibration and T—Test Spectra.

SOM classifier

Kohonen's self organising map (SOM) performs unsupervised classification through a two-layer architecture (see Figure 1), mapping input data into output classes (neurons). The input layer is a linear array with as many neurons as input variables (N, size of the input data vector). The output layer is a bidimensional matrix of processing elements and arbitrary size, K. Every output neuron is fully connected to all input nodes through a vector of connection weights. All weights are arranged in a $N \times K \times K$ matrix.

The classification process is divided into two sequential steps: training and runtime. During training, the network learns to identify patterns contained in the training set. Each data vector is presented to the network and activates an unique neuron, the so-called winner. The training algorithm will modify the winner's weight vector depending on its resemblance with the input pattern. This algorithm is not only applied to the winner neuron, but also to adjacent neurons (the neighbourhood, n) which also learn to weakly recognise the same pattern.

Pattern presentation is repeated while $\eta(t)$ (the time-varying learning parameter) and neighbourhood size are continuously decreased following a linear or exponential path. The process continues until n equals 1 and η reaches a predetermined final value. A standard option to enhance the network's performance is the use of reinforcements. A reinforcement consists for a sudden, small increase of the learning coefficient after the system has reached the final value for this

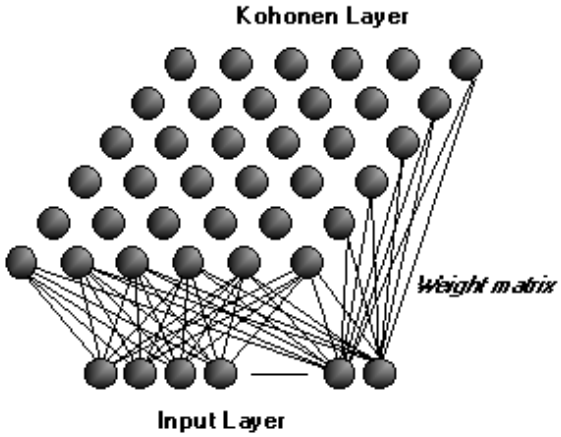

Figure 1. A self-organising map composed of two layers.

parameter. This is equivalent to introducing a random noise into the system or starting a new training cycle.

The classifier resulting from the training phase is described by a given weight matrix. At runtime, each node in the Kohonen layer will answer for a certain pattern previously learnt and will recognise all elements belonging to that class. This is achieved by presenting each vector once to the network and computing the winner neuron (the discriminant function is not modified). The self-organising training process guarantees that neighbour elements in the output layer share similar characteristics.

All experiments were carried out on a 486/DX4 100 MHz PC. Kohonen maps were defined and trained through a generic SOM-tool developed using Borland C++ under Windows 3.11. A set of pre- and post-processing programs for results analysis were also developed in C++.

Results

Several experiments have been carried out in order to determine the better neural architecture and training parameters. The resulting best classifier is obtained through a combination of reinforcement with linear $\eta(t)$ decreasing. Other initial conditions are: random initial weights (normalised with the Euclidean norm), η ranging between 0.7 and 0.01, 49 neurons in the Kohonen layer (square 7×7 matrix) and an initial neighbourhood equal to seven. Figure 2 depicts the Activation Map for the best classifier found through our experiments. The activation map is an snapshot of the Kohonen layer showing which neurons have recognised which patterns. Every substance is recognised by a single neuron except number 043, which is spliced into three subclasses. The spectra of this substance correspond to two different particle sizes (20 and 100 mess). Although the spectra are normalised, the network is able to separate different enough particle sizes. Characterisation of the best classifier requires a detailed analysis of both the dynamics of the learning process and the accuracy of the classification.

The dynamics of the learning process can be followed by generating a weight matrix every 50 presentations. Figure 3 shows the Activation Maps at 0, 200 and 1000 presentations. It can be seen how, at the beginning of the training process, all spectra are recognised by just one neuron [Figure

068		004		005		023
043 (100)		003		006		001
043 (100)						
043 (20)		012		021		069
064		028		033		013

Figure 2. Activation map for the best classifier. Each neuron shows the code of the substance recognised after training. Grey neurons are empty.

3(a)]. As the process progresses substances spread over the whole layer [Figure 3(b)] until the steady distribution is reached ([Figure 3(c)].

Application of the best classifier to fresh input data

Besides finding a model that correctly identifies all training substances, we want to evaluate the runtime performance of our classifier. This has been achieved by using 39 test spectra

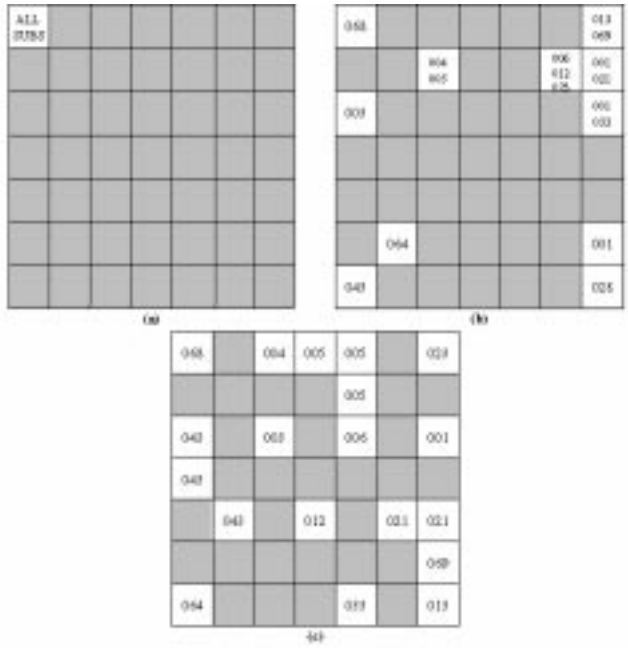

Figure 3. Activation Maps for (a) 0 presentations, (b) 200 presentations and (c) 1000 presentations.

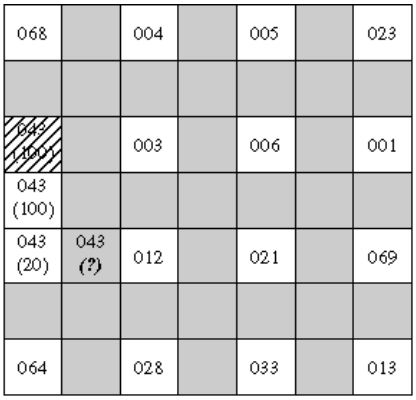

Figure 4. Activation map for the test set. Dashed neuron identified class 043 spectra only during training.

corresponding to the same raw substances measured under the same conditions. To classify each spectrum not included in the training process, the network computes the Euclidean distance between the normalised input vector and the weight matrix obtained through the best training process. The neuron showing the smallest distance is the winner. Thus, we say that the spectra belongs to this class. Since we are using the same substances for training and testing, a measurement of the classification accuracy is that spectra of test substances were classified by the same neuron as that the homologous training substance. Figure 4 shows the results obtained for the test set. All new data are correctly identified except for a single case: material 043. Two circumstances converge in this case:

- Training cases were divided into three adjacent neurons. As was discussed previously, this division partially corresponds to the existence of two different particle sizes. The classification process has correctly assigned a 100 mess test sample to the neuron that previously categorised this material. The same happens for the 20 mess test spectra.
- Two test spectra have been recognised by a previously empty neuron, adjacent to the 043 winners. Both correspond to the NIR analysis of a different manufacturer's sample. So, particle size (or other non-determined physical parameters) may unexpectedly vary. Nonetheless, The Kohonen network is able to identify the underlying similarity.

Conclusions

Non-supervised artificial neural networks have been revealed as an efficient method for substance identification because of their ability to construct non-linear mappings between high-dimensional variable spaces and low-dimensional class spaces without *a priori* information about the existing classes. Kohonen's self-organised map, although mathematically simple, implements a very powerful mechanism for input pattern categorisation.

A lot of experiments have been carried out in order to find a good classification model. Some ideas can be derived from the classification itself:

- When trying different Kohonen's layer sizes, smaller sizes lead to different substances being mixed. On the other hand, when a bigger size is used, classes are spliced. In other words, the bigger the Kohonen size, the higher the specialisation achieved, loosing in some way the concept of substance.

- The classification method makes no assumption on the classes size, neither guarantees a minimum class size. Thus, a balance has to be found between the number of classes and its size in order to keep classes homogeneous and big enough to allow the network to generalise. In the experiments shown here the better results have been obtained for size seven.
- Spectra normalisation with the Euclidean norm does not affect the ability of the network to differentiate between different particle sizes. So, it is possible to work with full raw data. There is no need for previous data reduction through other mathematical methods.

The existence of topological relationships between adjacent neurons would allow the finding of small changes in the shape of the spectrum as, for example, substance adulteration.

Finally, the behaviour of the network with a greater number of substances and worst repeatability conditions is still an open question. Nonetheless, our experiments reflect a real situation. Results are not laboratory test but a real life situation and we are confident of the advantages and generalisability of the method.

References

1. M.P. Derde and D.L. Massart, *Analytica Chimica Acta* **191,** 1 (1986).
2. T. Næs *et al., J. Near Infrared Spectroscopy* **1,** 1 (1993).
3. C. Borggaard *et al., Analytical Chemistry,* **64,** 545 (1992).
4. M.K. Alam *et al., Spectroscopy* **9,** 30 (1994).
5. U.M. Weigel and R. Herges, *J. Chem. Inf. Comput. Sci.* **32,** 723 (1992).
6. W.J. Jasper and E.T. Kovacs, *Textile Res. J.* **64,** 444 (1994).
7. T. Kohonen, *Self-Organization and Associative Memory*, 3rd Edn. Springer Verlag, New York (1989).
8. A.J. García-Tejedor *et al.*, "A Neural System for Short-term Load Forecasting Based on Day-type Classification". *Proceedings of Intelligent Systems Applied to Power Systems.* ISAP'95, pp. 353 (1994).