Chemometrics and Intelligent Laboratory Systems 149 (2015) 140–152

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

A non-standard view on artificial neural networks

Marjan Vračko *, Jure Zupan

National Institute of Chemistry, Ljubljana Slovenia

ARTICLE INFO

Article history:
Received 18 March 2015
Received in revised form 20 August 2015
Accepted 22 August 2015
Available online 30 August 2015

Keywords:
Error back propagation ANN
Counter propagation ANN
Bottle-neck
Decision hierarchy
Mem-computer elements
QSAR modeling
Classification
Fish toxicity

ABSTRACT

In the paper first the two main learning strategies of the artificial neural networks (ANNs), the error-back propagation (EBP) and Kohonen self-organizing maps (SOM) are briefly described. Next, two nonstandard network layouts of the ANNs (bottle-neck and pyramidal decision tree) one for each of both learning strategies are suggested. In the last part, the use of counter-propagation (CP) ANN for handling chemical structures in QSAR modeling, classification and clustering is discussed. These concepts are of particular interest for the computer-aided drug research and in computer-aided toxicology. In the case study the results of fish toxicity research are described.

1. Introduction

If a young scientist is acquainted with a new method he or she is usually prepared to try it immediately, just because it is new and fashionable. On the other hand, a more experienced chemist will be (as a rule) at least slightly suspicious when confronted with something new. Many years of experiences have taught us that the novel approaches although exciting and promising can after more intensive use and deeper insight show hidden flaws. This does not suggest that one has to avoid the new methods, it is rather a warning that a new method or approach should always be studied thoroughly and applied to several different situations with a great deal of scrutiny and attention before it can be adopted as a standard method. Not a single data handling method, artificial neural networks (ANNs) included, is applicable for all cases. Due to the fact that the ANNs are not one, but a group of methods, the attention and study of the condition with respect to desired outcome and the selection and quality of data involved should only be multiplied when one starts using them.

On the other hand one of the more attractive features of a new method is the possibility to look at the way to obtain the solutions from completely different perspectives. The present paper is focused on several situations in which the ANN methods can be employed in a rather nonstandard way.

2. The supervised and unsupervised methods

The broadest division of data handling methods runs along the line that separates the methods for visualization of data from the ones extracting quantitative values, determinations, rules, and decisions. This division overlaps very much (but not entirely) what is known in statistics, informatics and artificial intelligence as the separation between the unsupervised and supervised methods [1].

In the multi-variant problems, i.e., the problems involving not a single one, but m variables, the objects of interest (analyses, recipes, spectra, structures, chemical processes, etc.) are represented as m-dimensional vectors x:

$$x = (x_1, x_2, x_3, \ldots x_i, \ldots x_m). \quad (1)$$

Each object x, let us say an m-component mixture analysis of the sample x can be accompanied or supplemented with the measurements of n responses r, i.e., with p properties (physical, chemical pharmaceutical, and/or biological data) required by the quality control or any other

* Corresponding author.
E-mail address: marjan.vracko@ki.si (M. Vračko).

http://dx.doi.org/10.1016/j.chemolab.2015.08.016
0169-7439/

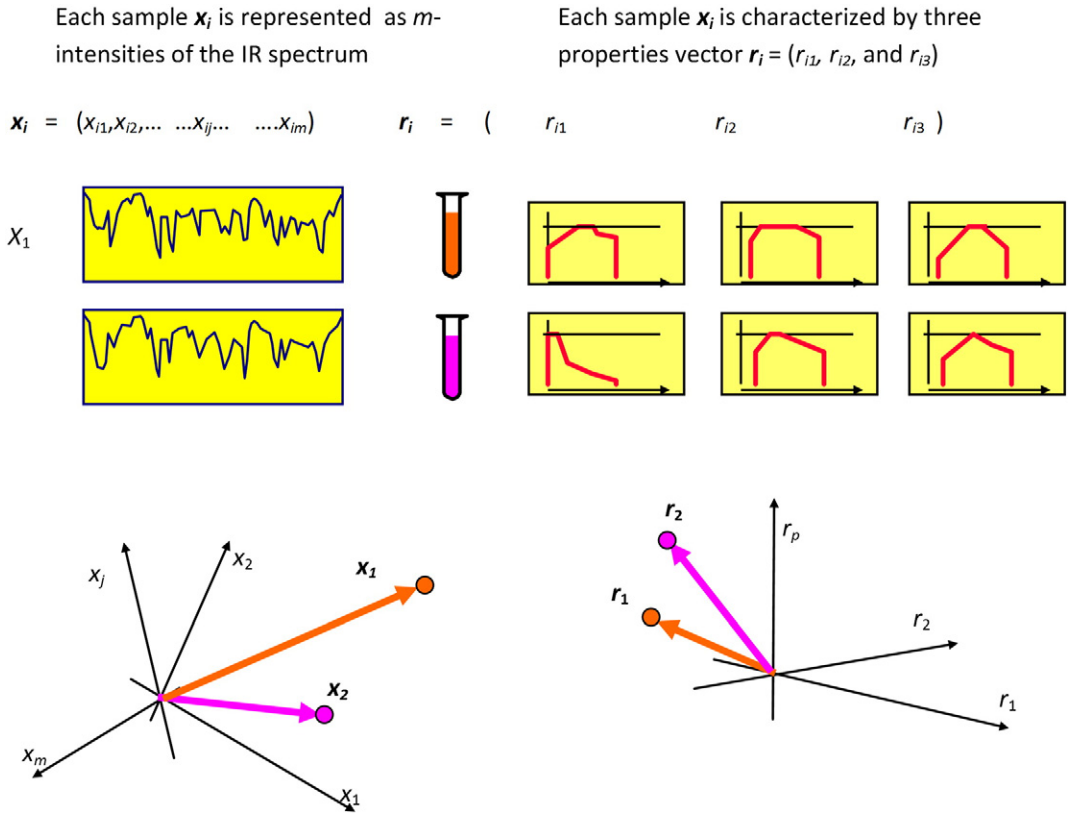

Fig. 1. Each of the two samples (objects) x_1 and x_2 are represented by m-intensities of their IR spectrum and associated with p properties or experimental responses written as p-dimensional vectors r_1 to r_p. The spectra x_i and the corresponding responses r_i can be shown as two vectors in two measurement spaces of different dimensions m and p, respectively.

standards of the particular product x (Fig. 1).

$$r = (r_1, r_2, r_3, \ldots r_j, \ldots r_p) \tag{2}$$

According to the availability of the response set $\{r_i\}$ complementary to the database $\{x_i\}$ two main types of data handling methods are used.

The methods that consider only the objects $\{x_i = (x_1, x_2 \ldots x_m)\}$, because the responses $\{r_i\}$ are not known or are withheld for whatever reason, are called the unsupervised. This means that the unsupervised methods are applied when one tries to find and describe the relative positions of the objects x_i and their mutual relations in the m-dimensional measurement space.

On the other hand, the supervised methods, are the ones that consider all pairs of the dataset $\{x_i, r_i\}$, i.e. the input (measurements) and the output (response set) data. Therefore, the supervised methods are used for finding (or for modeling) the ways to project the m-dimensional space defined by objects x_i into the p-dimensional space of responses r_i. Fig. 2 shows the relation between the two types of methods graphically.

There are various unsupervised and supervised methods for handling multivariate data, however in this paper we would like to focus our attention to the so called artificial neural network ANN techniques. As mentioned before, ANNs are not one, but rather a set of several methods ranging in their applicability from projection, clustering, and modeling to optimization. It is wise to keep in mind that in the phrase neural network the emphasis is on the word network rather than on the word neural. The meaning of this remark is to stress that the way artificial neurons are connected or networked together is far more important than the comparison between the action of biological and artificial neuron.

It is assumed that the readers are familiar with the basic items in the ANN vocabulary such as neurons, weights, inputs, outputs, targets, corrections, and learning parameters such as learning rate. The readers can find more detailed instructions about the ANN methodology in several books and papers [2,3]. In the following two paragraphs only a brief outline of the EBP and Kohonen networks will be given in order to assure consistent naming of the ANN features.

3. Error-back propagation ANNs

The most frequently applied ANNs are error-back propagation (EBP) [4] or perception ANNs [5] consisting of two or three layers of neurons. The name, error back-propagation comes from the fact that the ANN weights w^p_{ji} (connecting the input signal x_i with the output signal out_j on the layer p, (Eqs. (4), (5), and (7) in Box 1)) are adapted through the training by the feedback procedure, i.e., by correcting neuron weights starting at the neurons in the output layer and continuing towards the input. The change of the weights Δw^p_{ji} is calculated for any single neuron in the network each time an input vector $x_s = (x_{s1}, x_{s2}, \ldots, x_{sm})$ is sent through the ANN and its response vector $r_s = (r_{s1}, r_{s2}, \ldots, r_{sp})$ or modeled property r_s is compared to the last set of outputs $(out_{s1}^{last}, out_{s2}^{last}, \ldots, out_{sp}^{last})$ or to a single output of the network out_s^{last}, respectively (Eq. (4)). The corrections of weights in neurons of the last (output) layer are different from that of the rest (see Fig. 3). This is due to the fact that the actual error δ^p_j can be explicitly evaluated only at the output layer of the ANN (Eq. (4)) while δ^p for the hidden layers can only be estimated (Eq. (6)), because no actual target to which the hidden layer outputs out^k (k for hidden layers) could be compared exists. It can be said that the correction of weights flows from the output layer (where the error is estimated exactly) towards the weights in the input layer of neurons (where it is estimated by an average using Eq. (6)) in a backward-manner.