



ELSEVIER

Journal of Chromatography A, 825 (1998) 161–169

JOURNAL OF
CHROMATOGRAPHY A

Prediction of flame ionization detector response factors using an artificial neural network

M. Jalali-Heravi*, M.H. Fatemi

Sharif University of Technology, Department of Chemistry, P.O. Box 11365-9516, Tehran, Iran

Received 16 June 1998; received in revised form 14 August 1998; accepted 14 August 1998

Abstract

An artificial neural network (ANN) was successfully developed for the modeling of flame ionization detector response factors. The generated ANN was evaluated and applied for the prediction of response factors of several varieties of organic compounds. The results obtained using neural network were compared with different sets of experimental values as well as with those obtained using multiple linear regression technique. Comparison of neural network standard error of prediction values with those obtained using regression equations shows the superiority of ANNs over that of regression models. Calculations of Dietz response factor for two different prediction sets show that an ANN has a good predictive power. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Neural networks, artificial; Flame ionization detection; Response factors; Detection, GC; Regression models

1. Introduction

In the past decade, the topic of neural computing has generated widespread interest and popularity [1–3]. Neural computing is usually implemented by using artificial neural networks (ANNs). The popularity of this technique is due in part to the analogy between ANNs and biological neural networks. ANNs are thought to have the ability to learn during a training process, where they are presented with a sequence of inputs and outputs.

The use of ANNs in chemistry has grown substantially [4,5]. Zhang et al. reported the application of ANNs for modeling of fluorescence data [6]. Anker and Jurs have applied this technique for the prediction of the carbon-13 nuclear magnetic resonance (NMR) chemical shifts [7]. There are several reports

on the use of neural networks in the modeling of retention behavior and optimization of conditions in micellar liquid chromatography [8,9]. ANNs have been applied by several groups in quantitative structure–activity relationship (QSAR) studies [10,11]. These investigations concern the prediction of chromatographic retention index and selectivity coefficient of berberin selective electrode.

Applying gas chromatography (GC) as a tool for quantitative estimations of compounds requires knowledge of the response factor for each compound under the GC experimental conditions employed. Since numerous compounds are unavailable as standards, the development of a theoretical method for estimating response factor seems to be useful. The first work on prediction of flame ionization detector (FID) response factors (RFs) was published by Katritzky et al. in which they applied multivariate statistical partial least-square treatment [11]. They

*Corresponding author.

reported the use of the Dietz RF as a dependent variable and six structural and quantum-mechanical descriptors as independent variables in the prediction of FID RFs of some organic compounds [12].

The main aim of the present work was development of an ANN for the modeling of the FID RF. The generated ANN was evaluated and applied for the prediction of RFs of a wide series of organic compounds. The results obtained by this method were compared with different sets of experimental values as well as with those obtained using multiple linear regression (MLR) technique.

2. Methods

A detailed description of the theory behind a neural network has been adequately described elsewhere [13–17]. Therefore, only the points relevant to this work are described here. A fundamental procession element of an ANN is a node (Fig. 1). Each node has a series of weighted inputs, W_{ij} , and acts as a summing point of weighted input signals. The summed signals passes through a transfer function that may be in sigmoidal form. The output of node j , O_j , is given by Eq. (1):

$$O_j = 1/[1 + \exp(-X)] \quad (1)$$

where X is defined by the following equation:

$$X = \sum W_{ji} O_i + B_j \quad (2)$$

In Eq. (2), B_j is a bias term, O_i is the output of the

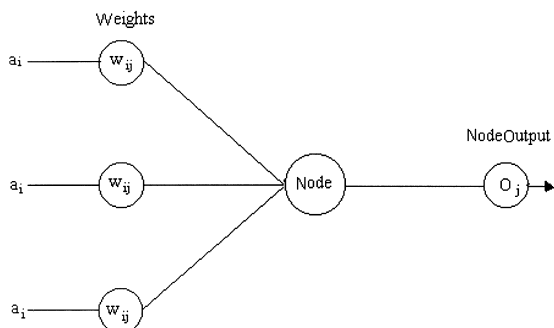


Fig. 1. Schematic representation of a node in an artificial neural network.

node of the previous layer and W_{ji} represents the weight between the nodes of i and j .

A feedforward neural network consisting of three layers is constructed as in Fig. 2. The first layer (input layer) consists of nodes and acts as an input buffer for the data. Signals introduced to the network, with one node per element in the sample data vector, pass through the input layer to the layer called the hidden layer. Each node in this layer sums the inputs and forwards them through a transfer function to the output layer. These signals are weighted and then pass to the output layer. In the output layer the processes of summing and transferring are repeated. The output of this layer now represents the calculated value for the node k of the network.

Training of backpropagation neural network, BNN, requires the comparison of the network output with an expected value. This comparison may be presented in an iterative fashion to the network with a weighted adjustment after each run. The differences between the output and the expected value backpropagated to the network and followed by adjustment of the weights and biases. The adjusted weights and biases can be calculated according to Eqs. (3) and (4).

$$\Delta W_{kj}(n) = \eta \delta_{pk} O_{pj} + \alpha \Delta W_{kj}(n-1) \quad (3)$$

$$\Delta B_{kj}(n) = \gamma \delta_{pk} O_{pj} \quad (4)$$

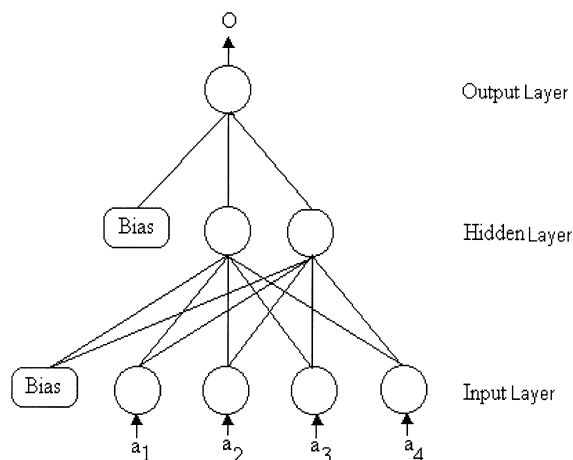


Fig. 2. Architecture for a three-layer back-propagation neural network.

In these equations ΔW_{kj} and ΔB_{kj} are the changes in the weights and biases between the node j in the hidden layer and the node k in the output layer, respectively; δ_{pk} is the error term obtained from the differences between the output and the expected value. The parameters η and γ in Eqs. (3) and (4) are learning rate of the weight and bias, respectively; α represents the momentum and n and $n-1$ refer to the present and the previous iterations, respectively.

Equations similar to the Eqs. (3) and (4) were used to adjust weights and biases connecting the hidden layers to the input one. The criterion for the stopping of the iteration during the training process could be a predefined number of iterations (p) or a desired difference between the output and its expected value. In order to obtain a parsimonious model, the network architecture was modified and tested. The number of hidden layer nodes, learning rates and momentum were parameters optimized.

In the present work for the sake of comparison a linear regression model was also developed and its results were compared with the calculated ANN response factors.

The main goal of this work was development of an ANN to predict the FID RF. Dietz relative response factors using *n*-heptane as a standard were considered throughout this work [12,18,19]. This factor depends on how many carbon atoms are being oxidized and in which manner the atoms of a molecule are combined.

3. Experimental

3.1. Data set

The data set was taken from Ref. [12]. This set was randomly divided into two groups, a training set (Table 1) and a prediction set (Table 2). The training set consists of 122 compounds and their RFs were used for network generation. The prediction set includes 24 compounds (set 1) and their RFs were used for evaluation of the generated network. In order to double check the predictive ability of the trained network we have taken another data set from Ref. [18]. Table 3 shows this set (set 2) which consists of 32 RFs of a variety of organic compounds.

For the training set and two prediction sets given in Tables 1–3, *n*-heptane acts as a reference with a defined RF of 1.00. The compounds studied in this work consist of hydrocarbons, halohydrocarbons, benzene derivatives, ethers, esters, alcohols, aldehydes, ketones and heterocyclic compounds.

3.2. Neural network generation

The ANN program was written in FORTRAN 77 in our laboratory. All of the calculations presented by the authors were carried out on a Hewlett-Packard 133 MHz pentium computer, Model HP Vectra VL. For each compound the quantum-mechanical descriptors were obtained using the MOPAC program (version 6) [20]. The number of input nodes in the NN was equal to the number of descriptors and the number of nodes in hidden layer were optimized. The initial weights were randomly selected from a uniform distribution that ranged between -0.3 and $+0.3$. The initial biases' values were set to be one. These values were optimized during training of the network. The input RFs were normalized by passing through a sigmoid function.

Before training, the network was optimized for the number of nodes in hidden layer, learning rate and momentum. In order to evaluate the performance of the NN, the mean square error (MSE), was used as a criterion. This parameter can be calculated by using the following equation:

$$\text{MSE} = 1/2 \sum (\text{RF}_{\text{cal}} - \text{RF}_{\text{exp}})^2 \quad (5)$$

The network was then trained using the training set by backpropagation strategy for optimization of the weights and biases. Training was stopped when MSE was less than or equal to 0.0067. It is worth noting that the difference between the normalized experimental and predicted values of RFs obtained by Katritzky et al. was 0.067 [12]. Therefore a criterion of one tenth of this value was chosen in the present work to improve the ability of the trained NN. The MSE converged to 0.0067 after 98 000 iterations.

3.3. Multiple linear regression model

The linear regression technique was based on the

Table 1
Training set and corresponding observed and ANN-calculated response factors

| No. | Compound | RF _{cal} | RF _{exp} | Δ | No. | Compound | RF _{cal} | RF _{exp} | Δ |
|-----|-------------------------------|-------------------|-------------------|----------|-----|------------------------------|-------------------|-------------------|----------|
| 1 | Heptane | 1.02 | 1.00 | 0.02 | 62 | Benzaldehyde | 0.80 | 0.81 | -0.01 |
| 2 | Dodecane | 0.93 | 0.96 | -0.03 | 63 | <i>trans</i> -Cinamaldehyde | 0.76 | 0.72 | 0.04 |
| 3 | Pentadecane | 0.87 | 0.90 | -0.03 | 64 | Benzonitrile | 0.81 | 0.91 | -0.10 |
| 4 | Cyclohexane | 1.04 | 1.04 | 0.00 | 65 | Benzyl cyanide | 0.81 | 0.81 | 0.00 |
| 5 | 1-Octane | 1.00 | 1.00 | 0.00 | 66 | 1-Bromodecane | 0.59 | 0.60 | -0.01 |
| 6 | 1-Decane | 0.97 | 0.94 | 0.03 | 67 | Chlorocyclohexane | 0.68 | 0.71 | -0.03 |
| 7 | Benzene | 1.04 | 1.09 | -0.05 | 68 | Iodobenzene | 0.39 | 0.43 | -0.04 |
| 8 | Toluene | 1.05 | 1.17 | -0.12 | 69 | <i>p</i> -Chlorotoluene | 0.76 | 0.78 | -0.02 |
| 9 | <i>o</i> -Xylene | 1.05 | 1.04 | 0.01 | 70 | <i>p</i> -Bromotoluene | 0.56 | 0.56 | 0.00 |
| 10 | Ethylbenzene | 1.06 | 0.93 | 0.13 | 71 | 1-Bromonaphthalene | 0.47 | 0.47 | 0.00 |
| 11 | Propylbenzene | 1.06 | 1.05 | 0.01 | 72 | Diphenyl ether | 0.80 | 0.86 | -0.06 |
| 12 | Mesitylene | 1.06 | 1.02 | 0.04 | 73 | Benzyl phenyl ether | 0.74 | 0.84 | -0.10 |
| 13 | 2-Methylstyrene | 1.01 | 1.15 | -0.14 | 74 | Butyl phenyl ether | 0.83 | 0.84 | -0.01 |
| 14 | Cyclohexylbenzene | 1.02 | 0.96 | 0.06 | 75 | 1,3,5-Trimethylbenzene | 0.47 | 0.46 | 0.01 |
| 15 | Hexamethylbenzene | 1.03 | 1.04 | -0.01 | 76 | Dioctyl sulfide | 0.58 | 0.55 | 0.03 |
| 16 | 4-Phenyl-1-butane | 1.03 | 1.08 | -0.05 | 77 | Diphenyl sulfide | 0.71 | 0.62 | 0.09 |
| 17 | Biphenyl | 0.91 | 0.90 | 0.01 | 78 | Phenyl disulfide | 0.63 | 0.54 | 0.09 |
| 18 | Indane | 1.04 | 1.13 | -0.08 | 79 | Benzylmethyl sulfide | 0.81 | 0.74 | 0.07 |
| 19 | Tetraline | 1.04 | 1.01 | 0.03 | 80 | 1-Methylpiperazine | 0.59 | 0.57 | 0.02 |
| 20 | <i>cis</i> -Stilbene | 0.88 | 0.76 | 0.12 | 81 | Quinoline | 0.78 | 0.82 | -0.04 |
| 21 | 1-Benzyl-naphthalene | 0.84 | 0.89 | -0.05 | 82 | Isoquinoline | 0.78 | 0.82 | -0.04 |
| 22 | 1,1'-Binaphthyl | 0.78 | 0.76 | 0.02 | 83 | 1,2,3,4-Tetrahydroquinoline | 0.82 | 0.85 | -0.03 |
| 23 | Phenanthrene | 0.78 | 0.79 | -0.01 | 84 | Indole | 0.70 | 0.61 | 0.09 |
| 24 | Anthracene | 0.80 | 0.76 | 0.04 | 85 | Benzimidazole | 0.64 | 0.49 | 0.15 |
| 25 | Fluorene | 0.82 | 0.86 | -0.04 | 86 | Phenanthridine | 0.66 | 0.66 | 0.00 |
| 26 | 1-Octanol | 0.70 | 0.77 | -0.07 | 87 | Acridine | 0.69 | 0.68 | 0.01 |
| 27 | 1-Dodecanol | 0.61 | 0.59 | 0.02 | 88 | 1-Methyl-2-pyridone | 0.48 | 0.48 | 0.00 |
| 28 | Cycloheptanol | 0.76 | 0.73 | 0.03 | 89 | Thiophene | 0.70 | 0.67 | 0.03 |
| 29 | Cyclohexanol | 0.74 | 0.80 | -0.06 | 90 | Pyrole | 0.76 | 0.74 | 0.02 |
| 30 | Phenol | 0.74 | 0.73 | 0.01 | 91 | Dibenzofurane | 0.73 | 0.66 | 0.07 |
| 31 | <i>p</i> -Cresol | 0.78 | 0.77 | 0.01 | 92 | Dibenzothiophene | 0.68 | 0.71 | 0.03 |
| 32 | <i>o</i> -Cresol | 0.79 | 0.82 | -0.03 | 93 | Benzothiazole | 0.63 | 0.63 | 0.00 |
| 33 | 2-Ethylphenol | 0.76 | 0.82 | -0.06 | 94 | 3-Picoline | 0.85 | 0.86 | -0.01 |
| 34 | 4-Ethylphenol | 0.77 | 0.78 | -0.01 | 95 | 4-Picoline | 0.85 | 0.86 | -0.01 |
| 35 | 4-Isopropylphenol | 0.76 | 0.72 | 0.04 | 96 | 2-Picoline | 0.85 | 0.88 | -0.03 |
| 36 | 2-Isopropoxyphenol | 0.64 | 0.72 | -0.08 | 97 | 2,3-Lutidine | 0.85 | 0.86 | -0.01 |
| 37 | Benzyl alcohol | 0.78 | 0.86 | -0.08 | 98 | 2,4-Lutidine | 0.85 | 0.86 | -0.01 |
| 38 | 1-Naphthol | 0.57 | 0.58 | -0.01 | 99 | 2,6-Lutidine | 0.86 | 0.86 | 0.00 |
| 39 | 5,6,7,8-Tetrahydro-1-naphthol | 0.71 | 0.82 | -0.11 | 100 | 2-Ethylpyridine | 0.85 | 0.80 | 0.05 |
| 40 | 1,3-Propanediol | 0.43 | 0.45 | -0.02 | 101 | 3-Ethylpyridine | 0.84 | 0.86 | -0.02 |
| 41 | 1,7-Dihydroxynaphthalene | 0.50 | 0.50 | 0.00 | 102 | 2,4,6-Collidine | 0.82 | 0.87 | -0.05 |
| 42 | Octylamine | 0.47 | 0.79 | -0.22 | 103 | 3-Cyanopyridine | 0.69 | 0.77 | -0.08 |
| 43 | Dodecylamine | 0.68 | 0.69 | -0.01 | 104 | 3-Pyridinecarboxaldehyde | 0.68 | 0.70 | -0.02 |
| 44 | Cyclohexylamine | 0.77 | 0.78 | -0.01 | 105 | 3-Acetylpyridine | 0.70 | 0.68 | 0.02 |
| 45 | Cycloheptylamine | 0.77 | 0.79 | -0.02 | 106 | 2-Amino-4,6-dimethylpyridine | 0.76 | 0.66 | 0.10 |
| 46 | Aniline | 0.86 | 0.82 | 0.04 | 107 | Ethyl pipercolinate | 0.52 | 0.53 | -0.01 |
| 47 | Benzylamine | 0.87 | 0.85 | 0.02 | 108 | Ethyl isonicotinate | 0.53 | 0.58 | -0.05 |
| 48 | <i>p</i> -Toluidine | 0.85 | 0.87 | -0.02 | 109 | 2,2'-Bipyridine | 0.69 | 0.73 | -0.04 |
| 49 | <i>N</i> -Methylaniline | 0.85 | 0.87 | -0.02 | 110 | Cyclohexanecarboxylic acid | 0.60 | 0.60 | 0.00 |
| 50 | Diphenylamine | 0.74 | 0.69 | 0.05 | 111 | Phenylacetic acid | 0.58 | 0.63 | -0.05 |
| 51 | Furfurylamine | 0.62 | 0.58 | 0.04 | 112 | 4-Phenylbutyric acid | 0.53 | 0.50 | 0.03 |
| 52 | 2,4-Dimethyl-3-pentanone | 0.78 | 0.78 | 0.00 | 113 | <i>trans</i> -Cinamic acid | 0.53 | 0.50 | 0.03 |
| 53 | Propiophenone | 0.79 | 0.84 | -0.05 | 114 | 2,4,5-Trimethylbenzoic acid | 0.59 | 0.57 | 0.02 |
| 54 | Cyclopentanone | 0.75 | 0.72 | 0.03 | 115 | Diethyl carbonate | 0.40 | 0.40 | 0.00 |
| 55 | Cyclohexanone | 0.75 | 0.76 | -0.01 | 116 | Benzyl acetate | 0.73 | 0.74 | -0.01 |
| 56 | Cyclohexyl phenyl ketone | 0.70 | 0.78 | -0.08 | 117 | Phenyl benzoate | 0.64 | 0.69 | -0.05 |
| 57 | Benzophenone | 0.64 | 0.60 | 0.04 | 118 | Dihydrocoumarine | 0.67 | 0.60 | 0.07 |
| 58 | Acetophenone | 0.76 | 0.81 | -0.05 | 119 | Isocoumarine | 0.81 | 0.82 | -0.01 |
| 59 | 1-Tetralone | 0.74 | 0.75 | -0.01 | 120 | Methylphenyl sulfoxide | 0.62 | 0.58 | 0.04 |
| 60 | 1,4-Naphthoquinone | 0.53 | 0.68 | -0.15 | 121 | Nitrobenzene | 0.61 | 0.63 | -0.02 |
| 61 | Benzil | 0.59 | 0.59 | 0.00 | 122 | <i>p</i> -Nitrotoluene | 0.60 | 0.66 | -0.06 |

Table 2

Prediction set 1 and corresponding observed and ANN-calculated response factors

| No. | Compound | RF _{cal} | RF _{exp} | Δ | No. | Compound | RF _{cal} | RF _{exp} | Δ |
|-----|-----------------------|-------------------|-------------------|----------|-----|------------------------------|-------------------|-------------------|----------|
| 1 | <i>n</i> -Decane | 0.97 | 0.94 | 0.03 | 13 | Cycloheptane | 0.75 | 0.78 | -0.03 |
| 2 | Allylbenzene | 0.99 | 1.04 | -0.05 | 14 | Bromobenzene | 0.54 | 0.55 | -0.01 |
| 3 | Cumene | 1.21 | 1.12 | 0.09 | 15 | Anisole | 0.86 | 0.79 | 0.07 |
| 4 | Diphenylmethane | 0.89 | 0.85 | 0.04 | 16 | Thioanisole | 0.79 | 0.78 | 0.01 |
| 5 | Naphthalene | 0.86 | 1.07 | -0.21 | 17 | 1,2,3,4-Tetrahydroisoquinone | 0.85 | 0.85 | 0.00 |
| 6 | Bibenzyl | 0.93 | 0.90 | 0.03 | 18 | Quinaldine | 0.80 | 0.77 | 0.03 |
| 7 | 1-Decanol | 0.72 | 0.70 | 0.02 | 19 | 4-Ethylpyridine | 0.85 | 0.82 | 0.03 |
| 8 | <i>m</i> -Cresol | 0.76 | 0.77 | -0.01 | 20 | 4-Cyanopyridine | 0.69 | 0.67 | 0.02 |
| 9 | 2-Isopropylphenol | 0.76 | 0.81 | -0.05 | 21 | 2-Acetylpyridine | 0.65 | 0.69 | -0.04 |
| 10 | 1,6-Dihydroxynaphthol | 0.56 | 0.58 | -0.02 | 22 | 4-(2-Aminoethyl)pyridine | 0.72 | 0.72 | 0.00 |
| 11 | Decylamine | 0.71 | 0.76 | -0.05 | 23 | 2-(2-Hydroxyethyl)pyridine | 0.63 | 0.60 | 0.03 |
| 12 | <i>o</i> -Toluidine | 0.87 | 0.86 | 0.01 | 24 | Coumarine | 0.61 | 0.65 | -0.04 |

Table 3

Prediction set 2 together with observed and calculated response factors using NN and regression model

| No. | Compound | RF _{reg} | RF _{NN} | RF _{exp} | A_{NN} | A_{reg} |
|-----|-----------------------------|-------------------|------------------|-------------------|----------|-----------|
| 1 | Butane | 1.12 | 1.04 | 1.09 | -0.05 | 0.03 |
| 2 | Pentane | 1.09 | 1.04 | 1.04 | 0.00 | 0.05 |
| 3 | Octane | 1.02 | 1.00 | 0.97 | 0.03 | 0.05 |
| 4 | 2,2-Dimethylbutane | 1.07 | 1.03 | 1.04 | -0.01 | 0.03 |
| 5 | 2-Methylpentane | 1.06 | 1.03 | 1.05 | -0.02 | 0.01 |
| 6 | 3-Ethylpentane | 1.04 | 1.02 | 1.02 | 0.00 | 0.02 |
| 7 | 4-Ethylheptane | 1.02 | 1.00 | 1.02 | -0.02 | 0.00 |
| 8 | 3,3-Dimethylheptane | 1.00 | 0.99 | 1.00 | -0.01 | 0.00 |
| 9 | 2,4,4-Trimethylheptane | 1.01 | 0.99 | 1.01 | -0.02 | 0.00 |
| 10 | 2,3,3,4-Tetramethylpentane | 1.00 | 0.99 | 0.99 | 0.00 | 0.01 |
| 11 | 1-Hexene | 1.03 | 1.03 | 0.99 | 0.04 | 0.04 |
| 12 | 1-Propanol | 0.67 | 0.64 | 0.60 | 0.04 | 0.07 |
| 13 | 1-Butanol | 0.70 | 0.69 | 0.66 | 0.03 | 0.04 |
| 14 | 2-Methyl-1-propanol | 0.70 | 0.69 | 0.68 | 0.01 | 0.02 |
| 15 | 2-Butanol | 0.70 | 0.69 | 0.63 | 0.06 | 0.07 |
| 16 | 1-Hexanol | 0.72 | 0.72 | 0.74 | -0.02 | -0.02 |
| 17 | Butanal | 0.78 | 0.71 | 0.62 | 0.09 | 0.16 |
| 18 | Octanal | 0.80 | 0.71 | 0.78 | -0.07 | 0.02 |
| 19 | Hexanoic acid | 0.57 | 0.57 | 0.63 | -0.06 | -0.06 |
| 20 | Octanoic acid | 0.74 | 0.68 | 0.65 | 0.03 | 0.09 |
| 21 | Methylcyclohexane | 1.03 | 1.02 | 1.01 | 0.01 | 0.02 |
| 22 | <i>m</i> -Xylene | 1.06 | 1.06 | 1.04 | 0.02 | 0.02 |
| 23 | 1-Methyl-3-ethylbenzene | 1.05 | 1.05 | 1.01 | 0.04 | 0.04 |
| 24 | 1,3,5-Trimethylbenzene | 1.05 | 1.06 | 0.98 | 0.08 | 0.07 |
| 25 | 1-Methyl-4-isopropylbenzene | 1.03 | 1.05 | 0.99 | 0.06 | 0.04 |
| 26 | <i>n</i> -Butylbenzene | 1.03 | 1.06 | 0.98 | 0.08 | 0.05 |
| 27 | Isobutyl acetate | 0.65 | 0.60 | 0.54 | 0.06 | 0.11 |
| 28 | 2-Butanone | 0.74 | 0.64 | 0.61 | 0.03 | 0.13 |
| 29 | 3-Heptanone | 0.81 | 0.73 | 0.71 | 0.02 | 0.10 |
| 30 | 3-Octanone | 0.81 | 0.72 | 0.80 | -0.08 | 0.01 |
| 31 | Isophorone | 0.86 | 0.84 | 0.85 | -0.01 | 0.01 |
| 32 | Isopentane | 0.93 | 0.92 | 0.91 | 0.01 | 0.02 |

Table 4
Specification of multiple linear regression model

| Descriptor ^a | Notation | Coefficient | Mean effect |
|---|----------|------------------|-------------|
| (1) Relative effective mass of carbon atoms ($m_c n_c / M_w$) | REMC | 0.966±0.068 | 0.617 |
| (2) Relative effective number of carbon atoms (n_c / n_a) | RENC | 1.018±0.129 | -0.364 |
| (3) Minimum valency of a hydrogen atom | MVHA | 0.672±0.119 | 0.608 |
| (4) Total hybridization component of the molecular dipole | THCMD | 0.009±0.006 | -0.014 |
| (5) Minimum total bond order (>0.1) of a carbon atom | MTBOC | 0.154±0.023 | -0.048 |
| (6) Total energy | TE | 0.000135±0.00002 | -0.210 |
| Constant | | 0.176±0.110 | |

^a n_a is the number of atoms in the molecule; m_c is the atomic mass of a carbon atom; n_c is the number of carbon atoms in the molecule which are connected only to other carbon and/or hydrogen atoms.

construction of linear mathematical model relating the observed response factors to numerically encoded structural descriptors. A total of nine separate molecular structure descriptors was calculated for each compound in the data set. Five of these descriptors were the same as those chosen by Katritzky et al. [12]. However, our descriptors also included core-core repulsion, ionization potential, heat of formation and total energy of the molecule. Linear models were formed by a stepwise addition of terms [21]. The best model obtained is given in Table 4. Comparison of our MLR model with that of Katritzky et al. indicates that the descriptor of “total molecular one center electron–electron repulsion energy” in their equation is replaced by the total energy of molecule in our model. However, the coefficients of descriptors in the present linear equation are completely different from those of Katritzky et al.’s model [12].

4. Results and discussion

In order to determine the optimum number of nodes in hidden layer several training sessions were conducted with different number of hidden nodes. The value of MSE was calculated and recorded after every 1000 cycles and for a total of 20 000 iterations. The calculated values of MSE were plotted against the number of iteration, from which the number of hidden nodes with minimum value of MSE was chosen (Fig. 3a). It can be seen from this figure that five nodes in hidden layer were sufficient for a good performance of the network. Learning rates of weight and bias and also momentum were

optimized in a similar way and the results are shown in Fig. 3b–d, respectively. It can be seen from Fig. 3 that the optimum values of the weight learning rate, bias learning rate and momentum are 0.99, 0.3 and 0.7, respectively. In addition it should be noted that the number of input and output nodes of the NN were 6 and 1, respectively.

To evaluate the neural network, the standard error of prediction (SEP) [22] of its results is compared with the SEP of the regression model in Table 5. The SEP was calculated using the previous regression model results for the molecules of the prediction set 1 [12] and included in Table 5. From comparison of the SEPs of two models one may conclude that both models are reliable in predicting of response factor. Comparison of the SEP values of two regression models in Table 5 reveals the superiority of our linear model over that of Katritzky et al.’s regression model in estimating of the prediction set 1 response factors. It should be noted that the coefficient of the REMC descriptor in our model is positive while that of the RENC is negative (Table 4). As can be seen from Table 4, the mean effects for the REMC and RENC descriptors are 0.617 and -0.364, respectively. The net effect of these descriptors is increasing of the response factor as the presence of carbon atoms increases in the molecule. This is in agreement with the experiment and in contrast to the previous work in which the coefficients of both descriptors are negative [12]. However, comparison of the neural network SEP values with those obtained using the regression equations reveals the superiority of the neural network over that of the regression models.

The predictive power of the trained NN was tested

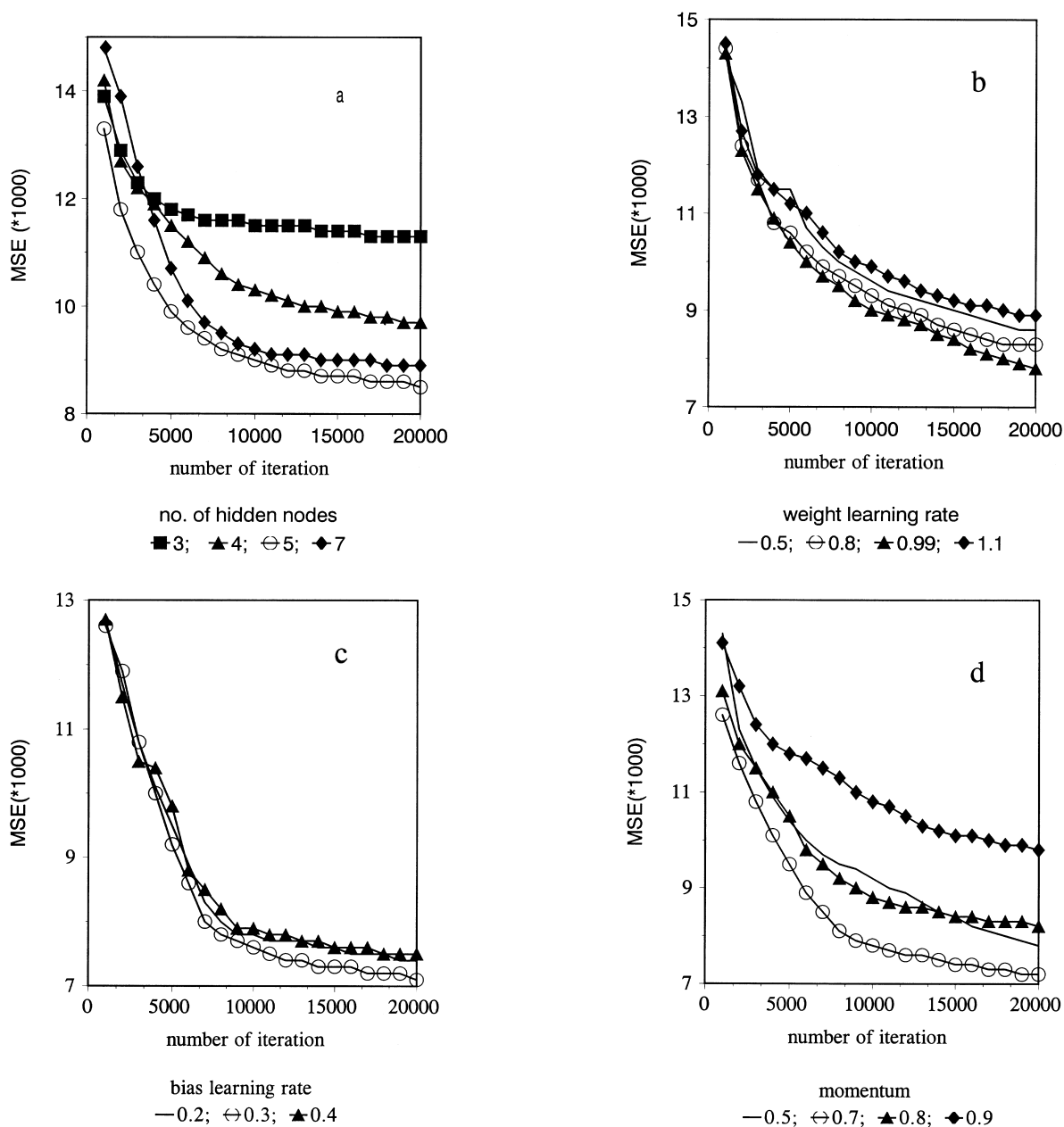


Fig. 3. Variations of MSE versus the number of iteration for optimization of (a) number of hidden nodes; (b) learning rate of weights; (c) learning rate of biases and (d) momentum.

by calculating the RFs of all molecules included in the data set. The experimental and NN calculated RFs for the molecules of training set and prediction set 1 are collected in Tables 1 and 2, respectively.

For comparison, the experimental and calculated RFs obtained by using the regression method are also included in Table 3 for prediction set 2.

Fig. 4 shows the plot of calculated against the

Table 5
SEP values for different sets studied in this work

| Data set | This work | | Previous work ^a |
|------------------|-----------|------|----------------------------|
| | NN | MLR | |
| Training set | 6.57 | 8.95 | 7.52 |
| Prediction set 1 | 7.12 | 5.37 | 7.49 |
| Prediction set 2 | 5.02 | 6.78 | |

^a SEPs calculated using the MLR model presented in Ref. [12].

experimental RFs for prediction sets 1 and 2, together with their correlation coefficients. The highest absolute error in the calculated RFs of the prediction sets was due to the naphthalene (Table 2). The calculated value of 0.862 for the RF of this molecule should be compared with the experimental value of 1.07. However, the calculated value in the previous work is 0.940 which is also lower than that of the observed value [12]. This indicates that the experimental value for the naphthalene may be over-estimated. The residuals of the NN predicted values of RFs are plotted against the experimental values in Fig. 5 after omitting this value. The propagation of residuals in both sides of zero indicate that no systematic error exists in the development of the neural network.

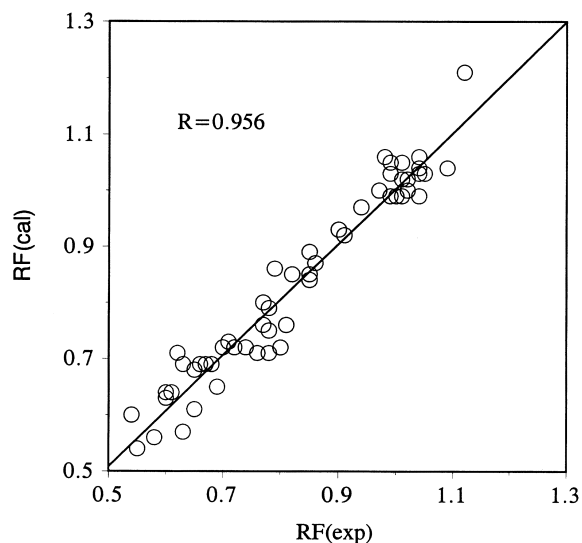


Fig. 4. Plot of calculated response factors against the experimental values.

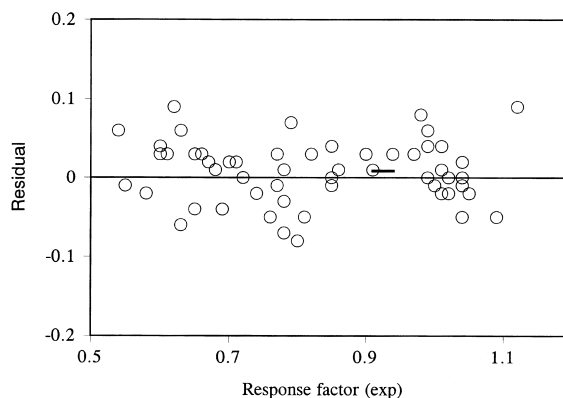


Fig. 5. Plot of residuals versus experimental factors.

Acknowledgements

The authors wish to acknowledge F. Parastaar for her helpful discussion.

References

- [1] P.K. Hopke, X. Song, *Anal. Chim. Acta* 348 (1997) 375.
- [2] H. Liu, X.W. Cao, R.J. Xu, N.Y. Chen, *Anal. Chim. Acta* 342 (1997) 223.
- [3] S.R. Junson, J.M. Sutter, H.L. Engelhart, P.C. Jurs, J. White, J.S. Kauer, *Anal. Chem.* 69 (1997) 4641.
- [4] N. Majcen, K. Rajer-Kanduc, M. Novic, J. Zupan, *Anal. Chem.* 67 (1995) 2154.
- [5] H. Chan, A. Butler, D.M. Falck, M.S. Freund, *Anal. Chem.* 69 (1997) 2373.
- [6] L. Zhang, J.H. Jiang, P. Liu, Y.Z. Liang, R.Q. Yu, *Anal. Chim. Acta* 344 (1997) 29.
- [7] S.L. Anker, C.P. Jurs, *Anal. Chem.* 64 (1992) 1157.
- [8] O. Jimenez, I. Benito, M.L. Marina, *Anal. Chim. Acta* 353 (1997) 367.
- [9] H.J. Metting, P.M.J. Coenegracht, *J. Chromatogr. A* 728 (1996) 47.
- [10] K.L. Peterson, *Anal. Chem.* 64 (1992) 379.
- [11] A.R. Katritzky, E.V. Gordeeva, *J. Chem. Inf. Comput. Sci.* 33 (1993) 835.
- [12] A.R. Katritzky, E.S. Ignatchenko, R.A. Barcock, V.S. Lobanov, M. Karelson, *Anal. Chem.* 66 (1994) 1799.
- [13] W.L. Xing, X.W. He, *Anal. Chim. Acta* 349 (1997) 283.
- [14] J.R. Long, V.G. Gregoriou, P.J. Gemperline, *Anal. Chem.* 62 (1990) 1791.
- [15] J.R. Long, H.T. Mayfield, M.V. Henley, P.R. Kroman, *Anal. Chem.* 63 (1991) 1256.
- [16] A.S. Pandya, R.B. Macy, in: *C++*, IEEE, CRC Press, Boca Raton, FL, 1996, Ch. 4.

- [17] R.A. Aliev, R.M. Vahidov, R.R. Aliev, *Artificial Neural Networks: Theory and Practice*, Tabriz University Press, Tabriz 1993, Ch. 5.
- [18] O.R. Anino, R. Villalobos, *Process Gas Chromatography*, Instrument Society of America, NC, 1992, Ch. 9.
- [19] R.L. Grob, *Modern Practice of Gas Chromatography*, Wiley, New York, 3rd ed., 1995, Ch. 5.
- [20] J.J.P. Stewart, MOPAC, A Semiempirical Molecular Orbital Program, QCPE 455(1983); Version 6 (1990).
- [21] N. Draper, H. Smith, *Applied Regression Analysis*, Wiley-Interscience, New York, 2nd ed., 1981, p. 307.
- [22] T.B. Blank, S.T. Brown, *Anal. Chem.* 65 (1993) 3084.