

Herbal drug raw materials differentiation by neural networks using non-metals content

Research Article

Bogdan Suchacz, Marek Wesolowski*

Department of Analytical Chemistry, Medical University of Gdansk,
80-416 Gdansk, Poland

Received 15 June 2010; Accepted 25 August 2010

Abstract: Three-layer artificial neural networks (ANN) capable of recognizing the type of raw material (herbs, leaves, flowers, fruits, roots or barks) using the non-metals (N, P, S, Cl, I, B) contents as inputs were designed. Two different types of feed-forward ANNs - multilayer perceptron (MLP) and radial basis function (RBF), best suited for solving classification problems, were used. Phosphorus, nitrogen, sulfur and boron were significant in recognition; chlorine and iodine did not contribute much to differentiation. A high recognition rate was observed for barks, fruits and herbs, while discrimination of herbs from leaves was less effective. MLP was more effective than RBF.

Keywords: Herbal raw materials • Artificial neural networks • Multilayer perceptron • Radial basis function network • Non-metals

© Versita Sp. z o.o.

1. Introduction

A high percentage of the world's population still depends on medicinal plants to treat many health problems. They are consumed, prepared as teas or infusions, and even applied directly to an injury without restriction in the false impression that natural products are always safe. Since they have not been tested in the same way as conventional pharmaceuticals, there is limited data confirming their therapeutic effectiveness and on adverse effects. The lack of standardization of herbal preparations and the inadequacy of consumer information is also of considerable concern [1]. Numerous cases of injury and even death caused by misuse, contamination or adulteration of medicinal herbs occur [2,3]. Therefore, their quality control is of particular importance, encompassing e.g. identity, purity, content, chemical and biological properties. Because herbal drugs are always mixtures of many constituents, developing criteria is much more complex and difficult. To verify the composition of herbal preparations, standards are required [3].

It was decided to explore the use of of plant bioelements to expand standardization procedures for medicinal plant raw materials. Six nonmetallic elements

- N, P, S, Cl, I and B were chosen because they are involved in all plant cell reactions. For example, nitrogen is in enzymatic proteins and nucleic acids, phosphorus is part of coenzymes and the energetic processes through ATP and ADP, sulfur is a component of enzymatic proteins, chlorine participates in water regulation, iodine is a constituent of some hormones, and boron modifies the cell membrane structure.

These non-metals are the components of secondary metabolites which determine the health care applications; there are close relations between the synthesis of biologically active compounds and accumulation of the elements in plants [4]. Their total content may indicate the levels of important substances from the biological point of view determining life functions and the proper development of medicinal plants and, in consequence, the quality of the herbal drugs obtained.

Due to fact that artificial neural networks are capable of modeling extremely complex nonlinear functions and can be used to discover relationships among great numbers of variables. The two most commonly-used types of ANNs – multilayer perceptron (MLP) and radial basis function (RBF) network - were chosen to reach the objective [5]. Six non-metals (nitrogen, phosphorus, sulfur, chlorine, iodine and boron) were used as the descriptors, and the task was to distinguish the herbal

* E-mail: marwes@gumed.edu.pl

part the sample represented, namely herbs, leaves, flowers, fruits, roots and barks. The herbal samples belonged to different species.

These networks were chosen because they handle classification problems in different manners - MLP uses hyperplanes to separate classes in pattern space, while RBF employs hyperspheres [5-7], therefore any existing relations in the concentration of non-metallic elements were to be recognized by either MLP or RBF network. A linear projection method – principal component analysis (PCA) was initially applied to identify patterns and outlying samples.

2. Experimental Procedure

2.1. Materials

Commercially available herbal raw materials - herbs (50 samples), leaves (27), flowers (29), fruits (16), roots (13) and barks (9), were included in the study. The description and detailed characteristics of the samples have been presented [8-12].

2.2. Determination

After mineralization the non-metals concentrations were determined by spectrophotometric methods, except sulfur. Sulfur was determined by turbidimetry using BaCl_2 . Details are given elsewhere [8-12].

2.3. Software

STATISTICA, (Statsoft Inc.) release 7.1 was used to create and train the artificial neural network models.

2.4. Calculations

To determine the type of herbal material using the non-metal content as inputs, two different types of artificial neural networks - MLP and RBF - were used. Their characteristics have been fully described elsewhere [5-7,13,14], so they will be briefly summarized.

Fig. 1 illustrates the networks' topology. They consisted of three layers: input, hidden, and output. All neurons in one layer were connected to those in the following one. The signals flowed from the input layer, advanced through hidden neurons to the output units, where the final result was obtained. Six (corresponding to the six non-metals) linear neurons were used as flow-through units in the input layer of both networks. The minimax algorithm scaled concentrations between the minimum and maximum of the data set [7]. In the output layer the number of neurons was determined by the number of groups into which the samples were classified.

In the MLP networks, a sigmoidal activation function was used in the hidden and output layer neurons together with a sum of squares error function. Radial basis function networks were composed of radial neurons with a Gaussian activation function in the hidden layer and linear units in the output layer. The number of neurons in the hidden layer was varied from 1 to 20.

Prior to the training process the overall data set was randomly divided into three subsets with 116 (80%) in training, 14 (10%) in validation and 14 (10%) in test sets. Each architecture with a different number of neurons in the hidden layer was trained using identical subsets, making the network performances comparable. To prevent the network from getting trapped in a local minimum, each was tested several times.

A two stage training process was applied to both types of ANNs. The MLPs were trained with the learning rate set to 0.1 and the momentum set to 0.2 using a back-propagation of error for 100 epochs and a conjugate gradient for 500 epochs. Training was halted when the validation error began to increase. In the training process six element target vectors had to be devised for six output neurons, so that the neuron representing each group of herbal samples generated 1, while the rest of the output neurons produced 0.

The standard training method was employed for RBF networks. In the first stage two algorithms were used. K-means assigned the radial centers in the data set and K-nearest neighbors calculated the distance from each center. The second stage optimized the output layer by a pseudo-inverse method.

Evaluation and comparison of the networks were based on two factors. The first was the root mean square (RMS) error of the training set, determined during training. When the RMS error became constant the training was stopped. It was calculated by:

$$RMS = \sqrt{\frac{\sum_{i=1}^n (y_i - t_i)^2}{n}}$$

where y_i was the value generated by the network, t_i was the target (desired) value, and n was the number of cases. The second factor was the proportion of cases in the test and validation sets correctly classified.

The best model of each type was chosen with the highest performance and the smallest RMS. The results were compared and the groups of herbal samples best recognized were identified.

Additional evaluation of the networks was done by examining the confusion matrix which provides a detailed breakdown of misclassifications. It illustrates the error tendencies of the network.

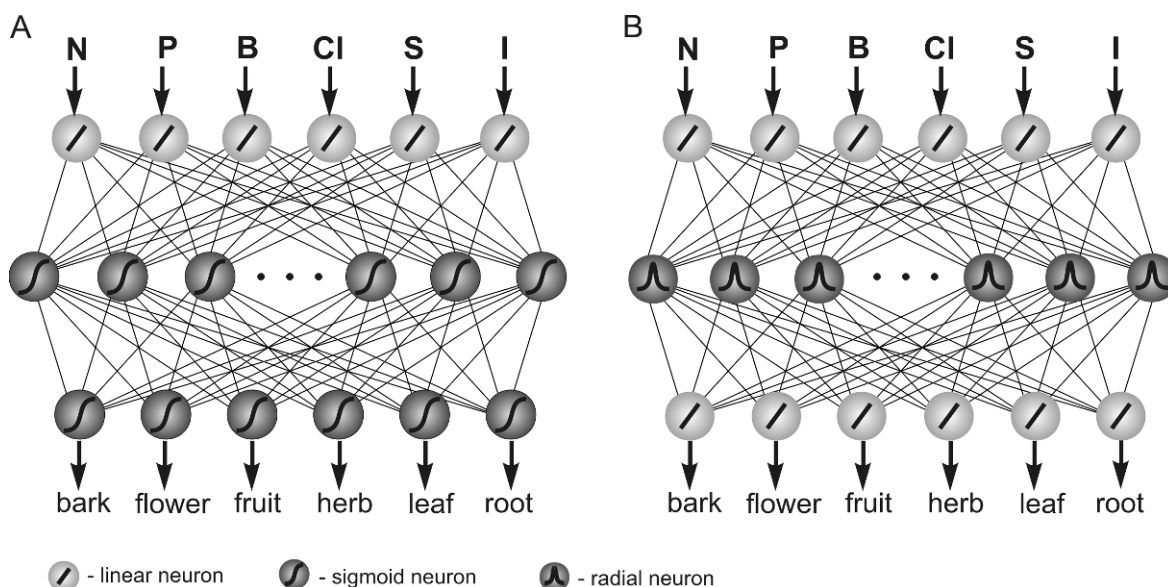


Figure 1. Network architectures: A) multilayer perceptron (MLP), B) radial basis function (RBF) network. The topology gives the number of neurons in the input, intermediate, and output layers.

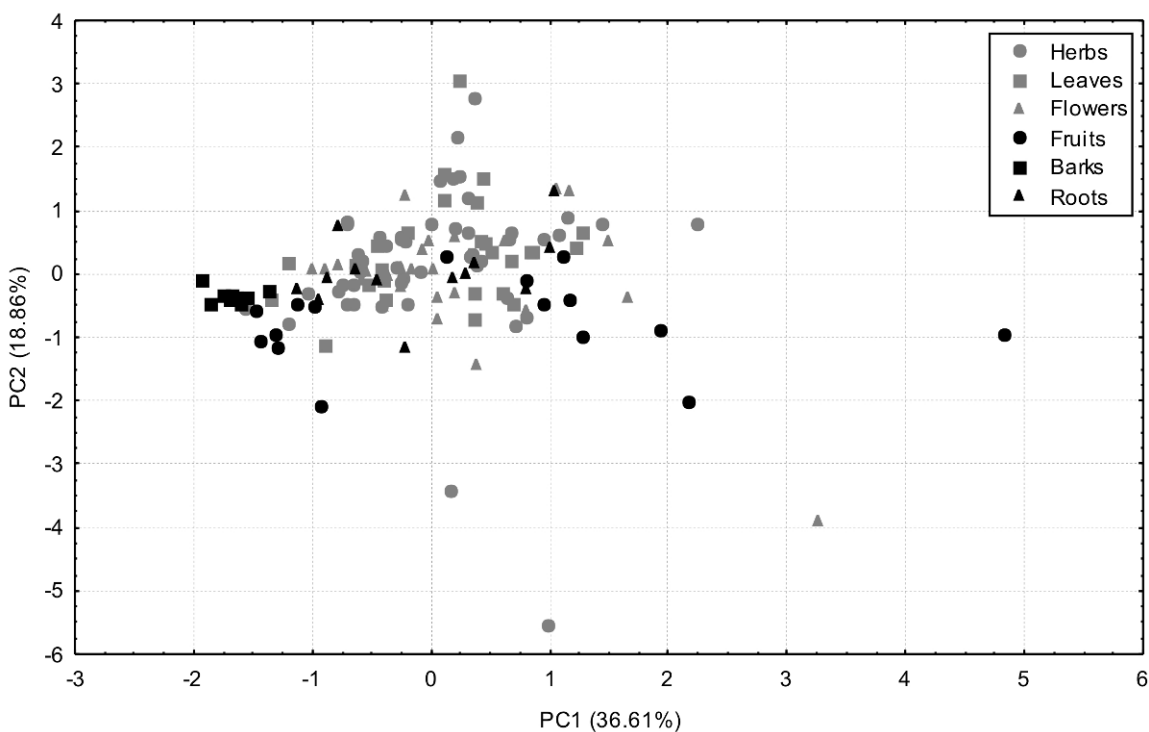


Figure 2. The first two principal component scores for herbal samples.

To identify the non-metals with the greatest impact on the recognition capacity of the neural networks the response surfaces were plotted. The complexity of the six-input response surfaces was reduced by simultaneously adjusting the concentrations of only two non-metals, holding the rest constant at zero. As a result each group of the samples was depicted by three simplified response surfaces. The impact of explanatory

variables (non-metals) on the response of the neural networks (recognition of herbal groups) was determined by the shape of the surfaces.

In regions where the slope of response to non-metallic concentration became steeper the network's response became stronger and, it correctly grouped the samples.

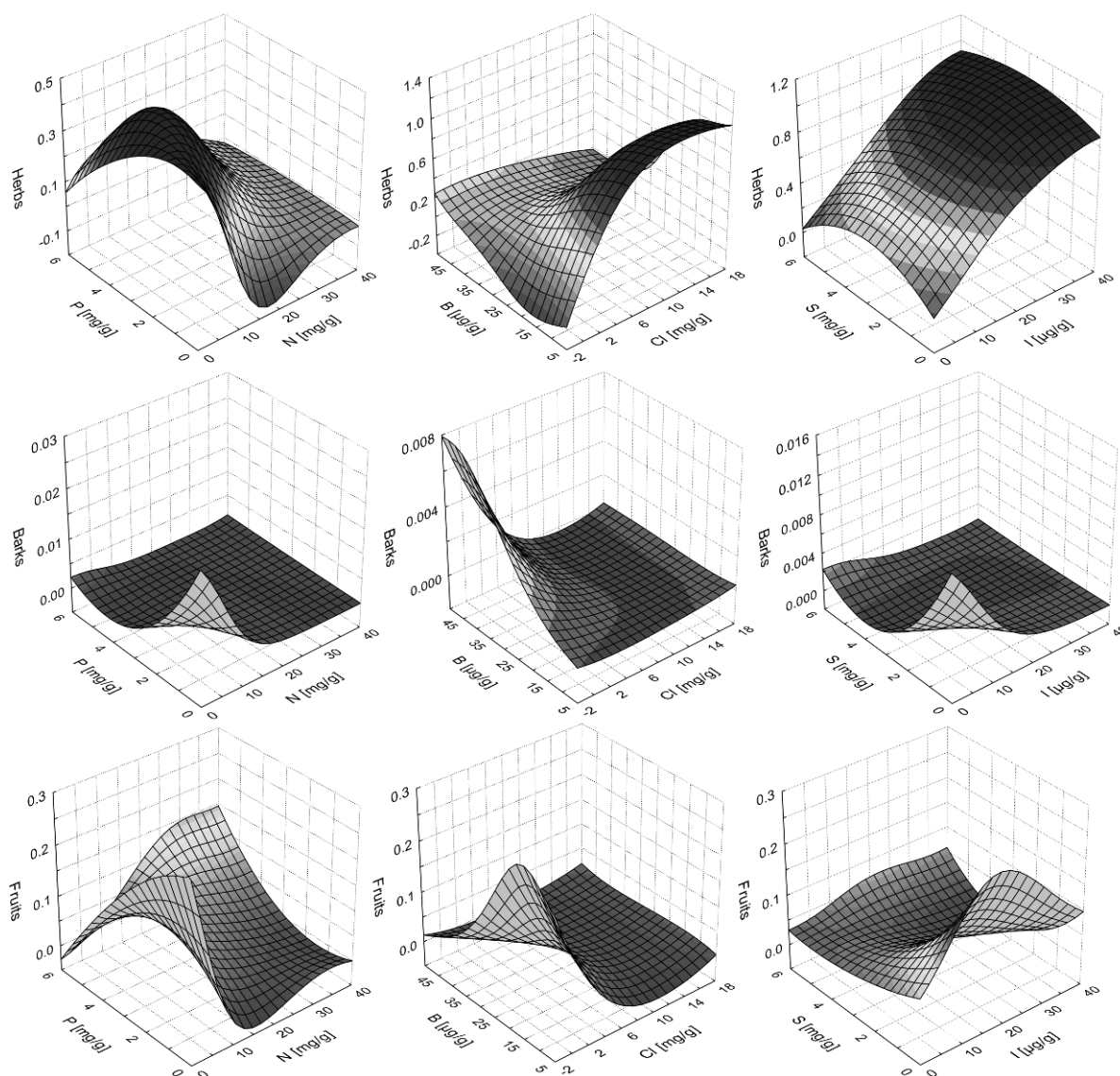


Figure 3. MLP response surfaces for herbs, barks and fruits.

3. Results and Discussion

3.1. Results of PCA analysis

Principal component analysis was first employed to preliminarily visualize differences among the classes and to identify outliers. As shown in Fig. 2, the first two principal components explain only 55.47% of the total variance and it is difficult to discern clear divisions among groups. Although the factor loadings (values greater than 0.7) of the first principal component (PC1) demonstrate the highest contribution of N, P and S and the second principal component (PC2) is connected with the Cl and I content, it is difficult to determine which non-metals differentiate the samples to the greatest extent,

because the samples are clustered in the center of the plot.

3.2. Results of ANN computations

The performance results and RMS rates for each ANN model are presented in Table 1. In the first column the network architectures are described by the number of neurons in the input, hidden, and output layers. The performance and error rates are listed for the training, validation and test subsets. The error rate is not directly interpretable in comparison with the performance measure, but is significant to the training algorithms – the lower its value, the better.

Summary statistics in Table 2 show that the samples best recognized by the MLP belonged to barks (89%),

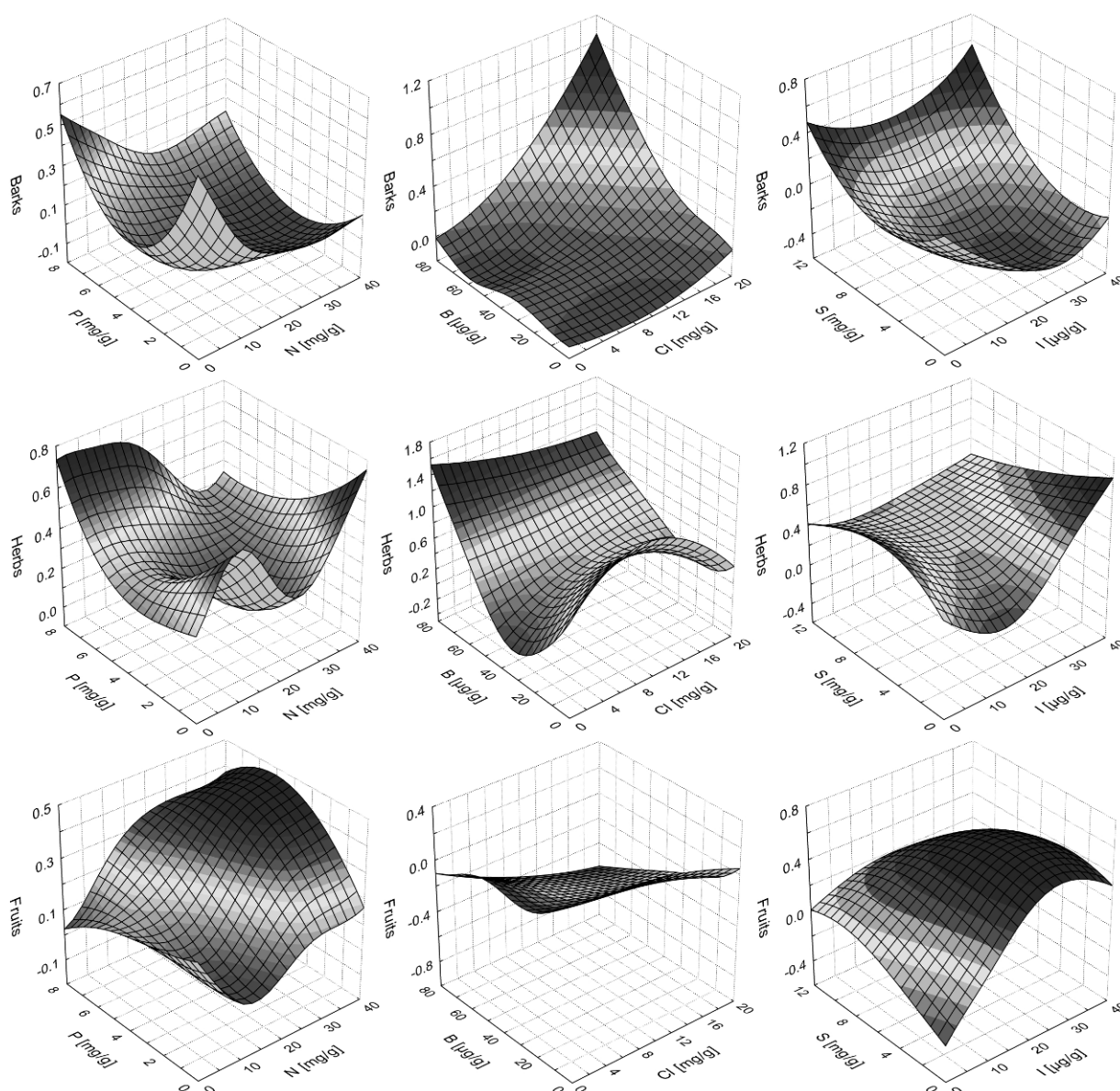


Figure 4. RBF response surfaces for barks, herbs and fruits.

Table 1. Performances and RMS values of the ANN models.

ANN Type Topology	Performance			Error Rate		
	Training	Validation	Test	Training	Validation	Test
MLP 6-14-6	0.7500	0.5278	0.5278	0.9947	1.5080	1.3626
RBF 6-12-6	0.6428	0.5667	0.5000	0.2952	0.3339	0.3474

herbs (78%), fruits (69%) and leaves (63%). The recognition rate for the flowers and roots was below 40%. The same samples were also best recognized by the RBF network. The artificial network recognized all bark samples, herbs (84%), fruits (almost 75%) and leaves (almost 60%). The group of roots was not identified at all

and less than 35% of flowers were identified. However, there were no samples unrecognized by both models.

The confusion matrix given in Table 3 shows that the MLP misclassified some herb samples but recognized most as leaves, and leaves as herbs. The worst classification occurred with flowers and roots, which

were also mostly identified as herbs. The RBF network similarly recognized herbs as leaves but fruits were labeled as barks.

The MLP response surfaces for the three best classified groups are in Fig. 3. For herb samples the elements most important in correct classification were N (15–20 mg g⁻¹) and P (1–5 mg g⁻¹). Boron's impact was observable for concentrations below 20 µg g⁻¹, while chlorine was significant above 10 mg g⁻¹. The content of I was most important between 30–40 µg g⁻¹. The barks' response surfaces indicated narrow ranges of the contents of analyzed non-metals influenced the classification, *i.e.*, N below 10 mg g⁻¹, P and S below 1 mg g⁻¹, Cl below 2 mg g⁻¹ and I below 5 µg g⁻¹. On the other hand, the fruits' identification the was based on the narrow range of B (20 to 30 µg g⁻¹), I (15–25 µg g⁻¹) and

S (< 2 mg g⁻¹). To a smaller extent the correct MLP output was affected by N (< 10 mg g⁻¹) and P (0.5–2 mg g⁻¹).

The RBF response surfaces for the best recognized groups are shown in Fig. 4. In the classification of barks, N (2 to 6 mg g⁻¹) and P (below 1 mg g⁻¹) were most important. Iodine (below 5 µg g⁻¹) also had an observable effect.

The RBF response surface in herb classification was more complex than that of MLP. Four non-metals were significant. The best classifiers were N (5–25 mg g⁻¹), P (< 2 mg g⁻¹), Cl (6–14 mg g⁻¹) and I (> 35 µg g⁻¹). B had no effect. In fruit identification the network mainly relied on P (practically the whole range 1–7 mg g⁻¹), S (2–8 mg g⁻¹) and N (> 20 mg g⁻¹). The remaining elements did not contribute substantially.

Table 2. Summary statistics for the applied architectures of ANNs.

ANN Type Topology	Raw Material	Correctly Classified (%)	Incorrectly Classified (%)
MLP 6-14-6	Barks	88.9	11.1
	Flowers	37.9	62.1
	Fruits	68.8	31.2
	Herbs	78.0	22.0
	Leaves	63.0	37.0
	Roots	38.5	61.5
RBF 6-12-6	Barks	100.0	0.0
	Flowers	34.5	65.5
	Fruits	74.9	25.1
	Herbs	84.0	16.0
	Leaves	59.3	40.7
	Roots	0.0	100.0

4. Conclusions

To some extent both types of artificial neural networks correctly recognized herbal raw materials based on non-metals content. In spite of the difference between the results for validation and test sets, rather high recognition of the herbal samples from the collection of barks, fruits and herbs was observed. There were no samples unrecognized by both ANN models.

Phosphorus and nitrogen are potentially valuable for standardization. Sulfur and boron have some significance. The concentrations of these four elements may be characteristic, particularly for the groups which were recognized most effectively. The concentration of chlorine and iodine did not contribute to the samples discrimination; thus they are not potential standardization parameters.

Table 3. Confusion matrices.

ANN Type Topology	Raw Material Recognized	Raw Material used as an input					
		Bark	Flower	Fruit	Herb	Leaf	Root
MLP 6-14-6	Bark	8	0	3	0	0	0
	Flower	0	12	0	3	2	3
	Fruit	1	0	11	0	1	0
	Herb	0	11	0	39	7	5
	Leaf	0	4	1	7	17	0
	Root	0	2	0	1	0	5
RBF 6-12-6	Bark	9	0	3	0	1	0
	Flower	0	10	1	3	1	3
	Fruit	0	5	12	0	1	2
	Herb	0	12	0	42	8	8
	Leaf	0	2	0	5	16	0
	Root	0	0	0	0	0	0

The most similarities were recognized in herbs and leaves. Samples of different groups were also mistakenly recognized as herbs. This can be explained by the fact that herb samples are composed of dried parts of whole plants including leaves, flowers and fruits in various proportions.

The performances of the two different types of ANN are comparable. However, the MLP network demonstrated a slightly better capability in identifying different herbal groups.

References

- [1] H.B. Matthews, G.W. Lucier, K.D. Fisher, *Environmental Health Perspectives* 107, 773 (1999)
- [2] P.R.G.M. De Smet, K. Keller, R. Hansel, R.F. Chandler, *Adverse Effects of Herbal Drugs* (Springer-Verlag, Heidelberg, 1997) vol. 1-3
- [3] W.M. Bandaranayake, In: I. Ahmad, F. Aqil, M. Owais (Eds.), *Modern Phytomedicine. Turning Medicinal Plants into Drugs* (Wiley-VCH, Weinheim, 2006) 25
- [4] M.Y. Lovkova, G.N. Buzuk, S.M. Sokolova, N.I. Kliment'eva, *Applied Biochemistry and Microbiology* 37, 229 (2001)
- [5] C. Bishop, *Neural networks for pattern recognition* (Oxford University Press, Oxford 1995)
- [6] S. Haykin, *Neural networks: A Comprehensive foundation*, 2nd edition (Prentice Hall, New Jersey, 1999)
- [7] *Statistica Neural Networks (manual)* (Statsoft Inc, Tulsa OK, 2007)
- [8] M. Wesolowski, P. Konieczynski, *Chemia Analityczna (Warsaw)* 41, 377 (1996)
- [9] M. Wesolowski, P. Konieczynski, B. Ulewicz, *Journal of Thermal Analysis and Calorimetry* 60, 299 (2000)
- [10] M. Wesolowski, P. Konieczynski, V. Medrzycka, *Chemia Analityczna (Warsaw)* 46, 697 (2001)
- [11] M. Wesolowski, P. Konieczynski, *Thermochimica Acta* 397, 171 (2003)
- [12] M. Wesolowski, P. Konieczynski, *International Journal of Pharmaceutics* 262, 29 (2003)
- [13] M. Wesolowski, B. Suchacz, *Fresenius Journal of Analytical Chemistry* 371, 323 (2001)
- [14] J. Zupan, J. Gastaiger, *Neural networks in chemistry and drug design* (Wiley, New York, 1999)