Available online at www.sciencedirect.com
SCIENCE DIRECT
Talanta 69 (2006) 37–42
Talanta
www.elsevier.com/locate/talanta

The recognition of similarities in trace elements content in medicinal plants using MLP and RBF neural networks

Bogdan Suchacz, Marek Wesołowski *

Department of Analytical Chemistry, Medical University of Gdansk, al. Gen. J. Hallera 107, 80-416 Gdansk, Poland

Received 27 April 2005; received in revised form 5 August 2005; accepted 10 August 2005

Abstract

The objective of the paper was to verify if the content of some elements provides enough information for proper classification of the medicinal plant raw materials. Such information could be helpful in standardization process of herbal products. Four elements—zinc, copper, lead and cadmium were determined using inverse voltammetry in commercially available medicinal herbal raw materials. Initially, principal component analysis (PCA) was employed to investigate the relationships among the analyzed trace elements. In the next stage of the study, two different types of feed-forward artificial neural networks (FANNs)—multilayer perceptron (MLP) and radial basis function (RBF) were applied. The concentrations of the elements were used as input variables to neural networks models, which were to recognize the taxonomy of the plant and the anatomical part it originated from. Although full recognition of the samples with use of FANNs on the basis of some trace elements content was not achieved, it was possible to identify two elements—cadmium and lead as the most important in the classification analysis of medicinal plants.

Keywords: Feed-forward artificial neural networks; Voltammetric determination; Classification; Inverse voltammetry; Principal component analysis

1. Introduction

In recent years there has been a considerable interest in herbal remedies and preparations which causes increasing demand for medicinal plant raw materials. Despite the serious degradation of the environment resulting from a dynamic development of industry in Europe and all over the world, the collection of plant material from natural habitats still remains the main source of plants commonly used in natural medicine [1].

Another source of raw materials for herbal industry is the cultivation of medicinal plants on specially designed plantations, where various plant species, which until recently were obtained exclusively from natural habitats have been grown [2].

Medicinal plants similarly to all living organisms are much diversified regarding the trace elements content [3–5]. Each element present in the soil, water or air can penetrate plant organs but the intensity of the intake increases proportionally to the concentration of the element in the environment around the plant. Generally speaking, trace mineral intake is dependent on plant's demand for a certain element, its availability in soil solution and different kinds of soil components.

Trace elements in plant tissues greatly affect pharmacotherapeutic properties of medicinal preparations obtained from the plants. As it is known plants synthesize organic compounds in photosynthesis, including those pharmacologically active, from various mineral components. While it is believed that the synthesis of organic compounds involves over 20 elements, there are many elements present in plant material whose physiological role is still largely unknown. At the same time it must be emphasized that all the elements play equally important role in the photosynthesis process [6]. When a plant lacks a certain element in the soil and its life process is disturbed, it is regarded as an essential component for a plant.

Considering all of the above and the fact that the concentration and spatial distribution of trace elements among soil, water, air and a plant itself maintain an almost con-

* Corresponding author.
E-mail address: marwes@amg.gda.pl (M. Wesołowski).

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.08.026

stant level in a certain geographical area, excluding sudden changes as a result of extreme weather conditions, i.e. floods or droughts [7], it was decided to study the possibility of making use of the information on trace elements content in plant raw materials used in health care, as a part of their identification in accordance with the anatomical part and the family to which the plant, being a source of the raw material belongs.

In the research, the four elements of anthropogenic origin such as: zinc, cadmium, copper and lead were included, for the reason that they are regarded as the contaminants of concern in assuring quality of herbal products [8,9]. After having determined the concentrations of the heavy metals in herbs commonly available in herbal stores and pharmacies in Poland, the primary intention was to identify relationships among them in analyzed dataset. The information about regularities existing among these potential contaminants would not only be helpful in standardization of herbal products [10], but would also draw the attention of manufacturers to more cautious analysis of raw plant materials characterized by increased abilities to accumulate these toxic heavy metals.

In order to achieve the objective, it was decided to employ two types of feed-forward artificial neural networks (FANNs), which are most commonly used in classification problems, namely multilayer perceptron (MLP) and radial basis function (RBF).

They were decided on because of their ability to detect complex non-linear relationships in the data, representing two different approaches to solving classification problems. MLP employs hyperplanes to divide the pattern space into various classes, while RBF uses hyperspheres.

Multilayer perceptrons trained with back-propagation algorithm have been well known to be capable of discovering relationships in datasets using information provided by suitable number of variables [11,12]. In this research, FANNs models were to recognize the taxonomy of the plant and its anatomical part on the basis of the concentrations of limited number of elements. In consequence, the secondary objective was chosen in order to confirm if RBF networks could offer more successful solutions operating on limited data (less time consuming, supplying additional information on existing relationships), thus indicating that the use of hyperspheres to divide up the pattern space into various classes is more advantageous when dealing with data described by a small number of variables.

2. Experimental

2.1. Material

The medicinal herbal raw materials commercially distributed in drug and herbal stores in Poland were included in the study. The collection of 318 samples represented five different anatomical parts of the plants—flowers (55), leaves (91), fruits (56), herbs (57) and roots (59).

From the above collection of the samples, the additional group was formed for the purpose of the taxonomical classification. As it was necessary for the plant families to be represented by a sufficient number of the samples, only the families with the number of samples >20 were chosen. As a result, 198 samples were included in the group representing the following six plant families—Apiaceae (21), Asteraceae (42), Ericaceae (20), Fabaceae (25), Lamiaceae (44) and Rosaceae (46). The figures in parentheses refer to the number of samples.

2.2. Sample preparation

Approximately 50 g samples of dry plant materials were homogenized at 20 °C for 20 s in a water-cooled grinder Knifetec 1095 (Foss Tecator, Höganäs, Sweden).

An accurately weighed 0.5 g portion of each sample was decomposed in 2 mL of nitric acid (65%, Selectipur, Merck) and 2 mL of hydrogen peroxide (30%, Suprapur, Merck) with the use of a high pressure microwave digestion system Uniclever BM-1z (Plazmatronika, Wroclaw, Poland). The process was run for 8 min at full power of magnetron (650 W) at programmed threshold pressure values (P_{max} = 45 atm, P_{min} = 40 atm). After the digestion procedure, the sample was placed in 25 mL volumetric flask and diluted with deionized water.

2.3. Voltammetric determination

Ten milliliters of deionized water, obtained by a triple distillation in Destamat Bi-18 system (Heraeus Quarzglas, Hanau, Germany), 1 mL of digestion solution and 2 mL of acetate buffer were added to the reaction vessel in the 757 VA Computrace (Metrohm, Herisau, Switzerland). The solution was degassed with nitrogen for 5 min to remove the electrochemically active oxygen. The determination of the trace elements—zinc, cadmium, lead and copper was carried out in a single voltammetric sweep at pH 4.6 with two standard additions using the hanging mercury drop electrode. The potentials were as follows: Zn (−0.97 V), Cd (−0.57 V), Pb (−0.38 V) and Cu (0.04 V).

The accuracy of the procedure was determined by the analysis of two certified reference materials (CRMs)—oriental tobacco leaves (CTA-OTL-1) and Virginia tobacco leaves (CTA-VTL-2) obtained from The Institute of Nuclear Chemistry and Technology, Warsaw, Poland. The level of the recovery ranged from 98.03 to 100.78% depending on the trace element.

2.4. Software

Statsoft Statistica release 6.0 was used to perform principal component analysis (PCA) and to construct and train the artificial neural network models.

2.5. Calculations

All information regarding the characteristics of PCA and FANNs employed in the study has been described in detail elsewhere [13–17], so it was decided to explain the applied procedure very briefly.

The dataset was divided into three subsets: training, validation and test so that the number of cases in a particular set amounts to 70–80% in training and 10–15% in validation and test, depending on the overall amount of cases in a given group.

The MLP and RBF networks applied in the study were composed of three layers: input, hidden and output. All units in one layer were connected to the units in the next one. The feed-forward flow of the signals proceeded from inputs, forwarded through hidden units and the final result was obtained in the output units.

The MLPs with sigmoidal activation function in the hidden and output layers and the sum of squares error function were trained for 300 epochs by a back-propagation of error algorithm. The training was stopped when the validation error began to increase. The learning rate was set to 0.5 and the momentum to 0.3. Radial basis function networks consisted of a linear input and output layers and a hidden layer of radial neurons modeling a Gaussian response surface. For RBF networks, a two-stage training process was applied. In the first stage two algorithms were used: *K*-means to assign the radial centers in the dataset and *K*-nearest neighbors to compute the deviation of each center. In the second stage the output layer was optimized with pseudo-inverse method.

Different architectures of networks were evaluated and compared on the basis of the value of the root mean square (RMS) error of the training dataset and the percentage of correctly classified (PCC) samples of test and validation datasets. In order to avoid getting trapped in a local minimum each developed network was tested six times. Every time different samples were randomly included in the training, validation and test subsets.

3. Results and discussion

In the preliminary study, PCA was used to examine the linear relationships among the analyzed elements and to determine if a reduced number of principal components (PCs) can be used to describe the raw dataset. The first two principal components with eigenvalues >1 are plotted in Fig. 1 separately for the samples representing the anatomical parts and the plant families. For the first group, the two PCs explain 65.15% of the total variance, while for the second group 75.08%. In both projections no evident separation between the samples has been observed. The contribution of the variables to the PCs is determined by the value of loadings. For the plot presented in Fig. 1A the loadings of the variables >0.7 were as follows: PC1—Zn, Cd and PC2—Cu. For the plot in Fig. 1B: PC1—Zn, Cd, Pb and PC2—Cu.

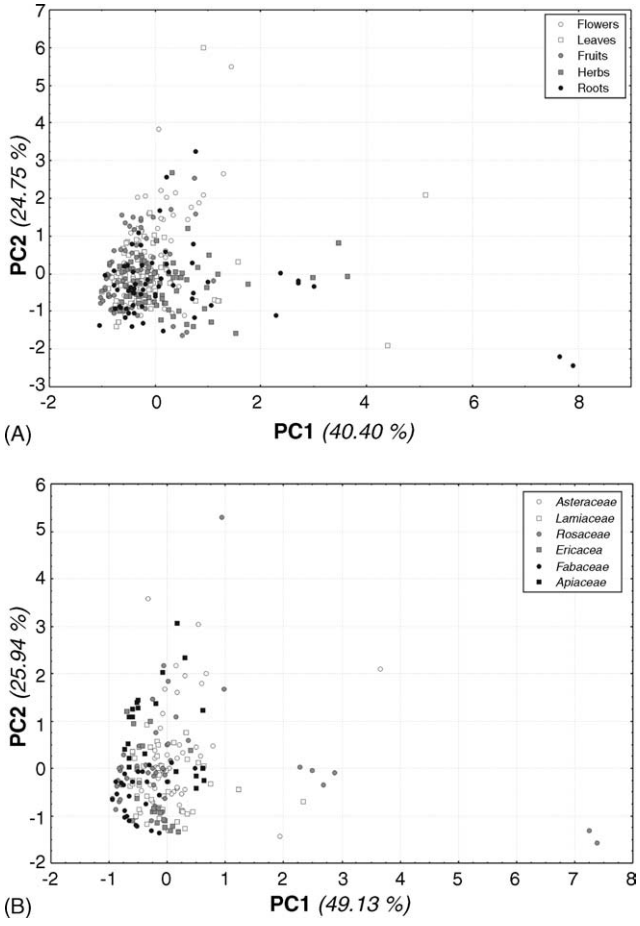

Fig. 1. A scatter plot of the first two principal components of the herbal samples according to (A) the anatomical part and (B) the plant family.

Even though it was easy to identify the levels of which trace elements were particularly associated with each component, the assignment of meaning to the PCs was difficult. Taking into consideration that the samples were located almost in one place and the values of explained variance, it was decided not to reduce the dimensionality from four to two, so that the actual structure of the data would not be obscured.

In further research, a neural network capable of classifying medicinal materials belonging to various plant families and a network that could recognize which anatomical part of a plant the sample represents were built. The concentrations of four heavy metals contents determined voltammetrically—zinc, lead, copper and cadmium were used as input variables to FANNs models, and the networks were to classify the herbal samples into six families: Apiaceae, Asteraceae, Ericaceae, Fabaceae, Lamiaceae and Rosaceae or five groups: flowers, leaves, fruits, herbs and roots. It must be noted, however, that the group of flowers included anthodia and inflorescences, while fruits were put together with seeds and roots with rhizomes.

In this research, it was not intended to use a large number of variables, because as it was shown in previous publications [18,19], in such cases the successful classification

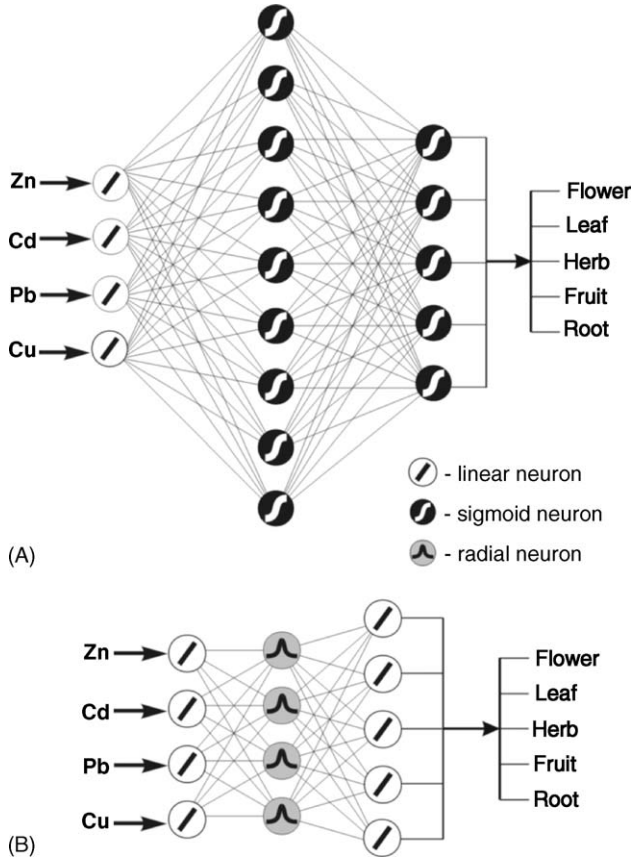

Fig. 2. FANNs models with the best performance used in the classification analysis according to the anatomical part. (A) Multilayer perceptron and (B) radial basis function network.

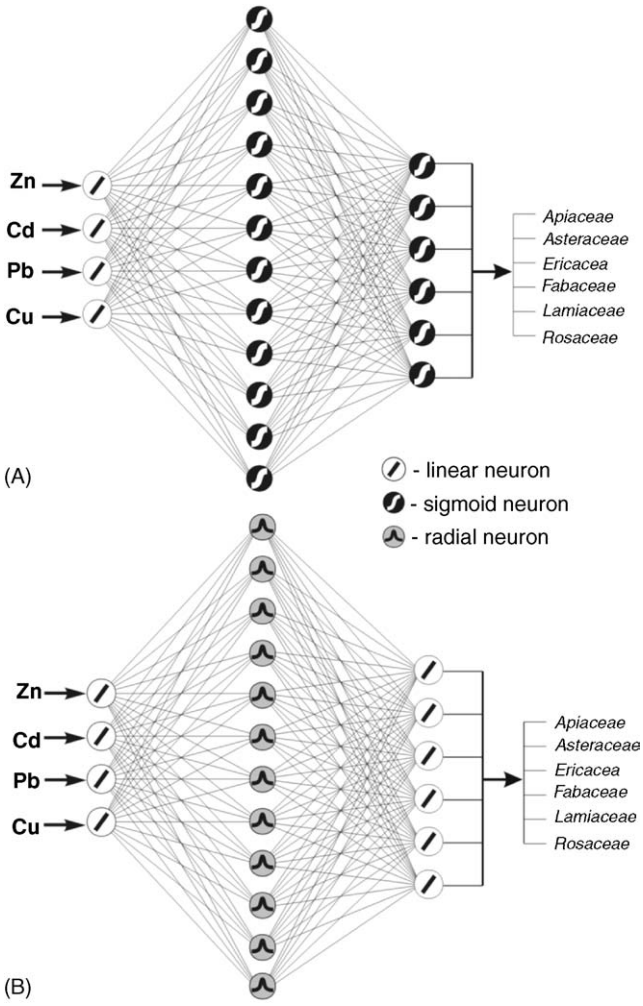

Fig. 3. FANNs models with the best performance used in the taxonomical classification. (A) Multilayer perceptron and (B) radial basis function network.

can be archived using solely principal component analysis. Therefore, working on the assumption that the correct determination of whether the sample belongs to a certain family or represents a specific part of a plant exclusively on the basis of the trace elements contents will be difficult to accomplish, it was decided to use two different types of artificial neural networks, particularly well suited for solving this kind of classification problems. Consequently any existing relationships in the analyzed dataset could be recognized by at least one of the employed FANNs.

From the group of network architectures, two models with the best performance (the highest PCC) and the smallest RMS, each representing different type were chosen. The architectures of FANNs employed in the study are illustrated in Figs. 2 and 3, whereas the classification results and the values of RMS for each FANN model are presented in Tables 1 and 2.

The network which was able to recognize the plant samples in the most effective way turned out to be the MLP in both classifications. The samples correctly recognized in the highest percentage which was about 70% belonged to the following groups: flowers, fruits and leaves. In comparison, the RBF networks were characterized by a weak ability to identify samples with the exception for leaves for which the PCC was >60%. Some herbs were identified as flowers or leaves by both FANNs types. This identification can be explained by the fact that herb samples consist of dried parts of whole plants including leaves, flowers, fruits and non-ligneous stems in various proportions.

In classification according to the plant family the best results were achieved using the MLP for the samples belonging to Apiaceae, Lamiaceae and Asteraceae with the value of PCC between 73 and 85%. The RBF network was also able to classify the samples from the above mentioned families in a highest degree, but the PCC was lower, except for Asteraceae family for which it amounted to about 91%.

A high value of the RMS errors and the comparison of the results for validation and test sets indicate a lack of generalization ability for any FANN model. The number of correctly recognized samples in each set differs greatly, ranging in most cases from 0 to 50%. However particular attention should be given to the fact, that there were no unrecognized samples in both classifications.

Table 1
The classification results of medicinal plant samples using MLP and RBF networks, with regard to the anatomical part they represent, on the basis of the trace elements content

Network type RMS	Samples	Training set		Validation set		Test set	
		Correctly classified (%)	Misclassified (%)	Correctly classified (%)	Misclassified (%)	Correctly classified (%)	Misclassified (%)
MLP 0.2872	Flowers	69.57	30.43	66.67	33.33	33.33	66.67
	Leaves	72.06	27.94	77.78	22.22	44.44	55.56
	Fruits	69.39	30.61	100	0	33.33	66.67
	Herbs	49.06	50.94	57.14	42.86	57.14	42.86
	Roots	48.98	51.02	66.67	33.33	66.67	33.33
RBF 0.3867	Flowers	43.48	56.52	66.67	33.33	16.67	83.33
	Leaves	67.65	32.35	77.78	22.22	55.56	44.44
	Fruits	0	100	0	100	0	100
	Herbs	56.60	43.40	57.14	42.86	71.43	28.57
	Roots	4.08	95.92	0	100	0	100

Table 2
The classification results of herbal raw materials using MLP and RBF networks, with regard to the plant family they belong to, on the basis of the trace elements content

Network type RMS	Samples	Training set		Validation set		Test set	
		Correctly classified (%)	Misclassified (%)	Correctly classified (%)	Misclassified (%)	Correctly classified (%)	Misclassified (%)
MLP 0.2894	Apiaceae	73.33	26.67	66.67	33.33	33.33	66.67
	Asteraceae	85.29	14.71	100	0	75.00	25.00
	Ericaceae	42.86	57.14	0	100	33.33	66.67
	Fabaceae	31.58	68.42	100	0	0	100
	Lamiaceae	75.00	25.00	100	0	50.00	50.00
	Rosaceae	55.26	44.74	50	50	50.00	50.00
RBF 0.3209	Apiaceae	66.67	33.33	66.67	33.33	33.33	66.67
	Asteraceae	91.18	8.82	75.00	25.00	50.00	50.00
	Ericaceae	28.57	71.43	0	100	33.33	66.67
	Fabaceae	31.58	68.42	100	0	0	100
	Lamiaceae	63.89	36.11	100	0	50.00	50.00
	Rosaceae	52.63	47.37	50.00	50.00	50.00	50.00

In addition, the sensitivity analysis was conducted to provide some information about the relative significance of each element concentration for the proper classification. In classification regarding the anatomical parts two FANNs models pointed to cadmium as the most important variable, while in the taxonomical classification lead was indicated.

4. Conclusion

In PCA analysis the variables responsible for the largest variation in the data were indicated, but no reduction of dimensionality was possible due to a lack of apparent division between the samples in both projections of the data.

Poor correlation between the results for validation and test sets for both MLP and RBF in the classification analysis of medicinal plant raw materials indicates that they are not able to fully recognize the family or anatomical part from which the samples originate, exclusively on the basis of zinc, copper, lead and cadmium content. Although the determination of those four elements does not provide sufficient information for the proper classification, high recognition of the herbal samples from three families: Apiaceae, Lamiaceae and Asteraceae and from the collection of flowers, fruits and leaves was observed. It may suggest that the content of the elements is in some way characteristic for these groups.

In addition, it must be stated that even though the FANNs models were operating on limited data, the MLP was indicated as the neural network characterized by a better ability to recognize medicinal plants with respect to the heavy metals concentrations. Two elements specified as the most significant classifiers were cadmium and lead.

References

[1] J. Sumner, The Natural History of Medicinal Plants, Timber Press, Portland, Oregon, 2000.
[2] WHO, WHO Guidelines on Good Agricultural and Collection Practices (GACP) for Medicinal Plants, WHO, Geneva, 2003.

[3] E. Epstein, A.J. Bloom, Mineral Nutrition of Plants: Principles and Perspectives, second ed., Sinauer Associates, Sunderland, Massachusetts, 2004.
[4] M.L. Berrow, J.C. Burridge, in: E. Merian (Ed.), Metals and their Compounds in the Environment, Occurrence, Analysis and Biological Relevance, VCH, Weinheim, 1991, pp. 399–410.
[5] L.V. Kochian, in: B.B. Buchanan, W. Gruissem, R.L. Jones (Eds.), Biochemistry and Molecular Biology of Plants, American Society of Plant Physiologists, Rockville, Maryland, 2001, pp. 1204–1249.
[6] F.B. Salisbury, C.W. Ross, Plant Physiology, fourth ed., Wadsworth Publishing, Belmont, California, 1992.
[7] A. Kabata-Pendias, H. Pendias, Trace Elements in Soil and Plants, third ed., CRC Press, Boca Raton, Florida, 2000.
[8] A.A.K. Abou-Arab, M.S. Kawther, M.E. El Tantawy, R.I. Badeaa, N. Khayria, Food Chem. 67 (1999) 357.
[9] A.M.O. Ajasa, M.O. Bello, A.O. Ibrahim, I.A. Ogunwande, N.O. Olawore, Food Chem. 85 (2004) 67.
[10] M.R. Gomez, S. Cerutti, R.A. Olsina, M.F. Silva, L.D. Martinez, J. Pharm. Biomed. Anal. 34 (2004) 569.
[11] J. Zupan, J. Gastaiger, Neural Networks in Chemistry and Drug Design, Wiley, New York, 1999.
[12] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.
[13] I.M. Farnham, K.H. Johannesson, A.K. Singh, V.F. Hodge, K.J. Stetzenbach, Anal. Chim. Acta 490 (2003) 123.
[14] M. Wesołowski, B. Suchacz, Fresenius J. Anal. Chem. 371 (2001) 323.
[15] M. Wesołowski, B. Suchacz, J. Therm. Anal. Calorim. 68 (2002) 893.
[16] M. Wesołowski, B. Suchacz, P. Konieczyński, Comb. Chem. High Throughput Screen. 6 (2003) 811.
[17] S. Haykin, Neural Networks: A Comprehensive Foundation, second ed., Prentice Hall, New Jersey, 1999.
[18] M. Wesołowski, P. Konieczyński, Int. J. Pharm. 262 (2003) 29.
[19] M. Wesołowski, P. Konieczyński, Thermochim. Acta 397 (2003) 171.