Artificial Neural Network for Identification of Multicomponent Mixtures of Tea

Mariyana Sestrimska[1], Tanya Titova[2] and Veselin Nachev[3]

Abstract – **In the present paper two types of artificial neural networks adapted to recognize the multicomponent tea mixed with fruit ingredients are presented. The experimental study is based on VIS/NIR spectroscopy, measurement of color and pH. Principal components analysis (PCA) is use to reduced high dimensional feature space to three specific factors.**

The results of neural networks: Backpropagation Artificial Neural Network (BP-ANN) and Kohonen's Self-organizing map (SOM) shows recognize the analyzed samples with an accuracy of 99.4% and 98%, respectively.

Keywords – **VIS/NIR spectral analysis, principle component analysis, artificial neural network, self-organizing maps, tea**

I. Introduction

Tea is a beverage which is obtained by brewing of various herbs, dried flowers or fruits of various plants. It is among the most consumed beverage in the world.

The qualities of tea are determined by a large content of various organic compounds such as polyphenols, alkaloids, flavonoids, caffeine and others. The leaves of the tea plant contains many vitamins such as vitamin C, E, B1, B6, carotene, folic acid, and also the minerals - manganese, potassium, fluoride and etc. [1,2].

Traditionally, the quality characteristics of tea are determined by a combination of organoleptic analysis and conventional analytical instrumentation. These methods are costly in terms of the time and labor, and also inaccurate for a number of reasons.

The purpose of this paper is to synthesize and train artificial neural networks, in particular Backpropagation Artificial Neural Network (BP-ANN) and Kohonen's Self-organizing map (SOM), to identify the types of tea with different fruit ingredients.

II. Materials and Methods

A. Preparation of the Samples

This study used 500 samples of six different types of tea multicomponent mixtures purchased from market. Samples are pre-selected in six groups, each of which receives an identification code (label). The distribution of the tea samples in groups, and the label of each of the groups are presented in Table I.

TABLE I
GROUPING OF TEA SAMPLES

Identification code (label)	Number of samples	Ingredients
Cluster 1	50	Chamomile, hibiscus, walnuts, almonds Dried fruits: papaya, passion fruit, melon, bigaradiya
Cluster 2	60	Ceylon mélange tea
Cluster 3	60	Green tea (Camellia sinensis), ginkgo biloba, lemon grass
Cluster 4	60	Ceylon tea, lemon
Cluster 5	120	Black tea, rose leaves (10%), rose flavoring (3%)
Cluster 6	150	Tea (Camellia sinensis), strawberries, raspberries, blackberries, blueberries

** Information on ingredients and origin of the tea samples is taken from the product label.*

The tea packages (2-3g) was put (from 3 to 5 minutes) in the distilled water which is heated at 100°C previously.

So prepared samples were cooled to the room temperature 22-23°C, poured in 20 ml glass tubes closed and prepared for further processing.

All samples of tea are prepared and stored under identical conditions.

B. Measurement on Color and pH

To determine the pH of the samples tested tea used digital pH-meter PH-201, the measurement was performed at the room temperature.

Color characteristics of tea samples (L*, a*, b*, c*, h*) were measured by a spectrophotometer optic USB4000, equipped with a suitable sample holder and the necessary optics. In CIELAB color space, chromaticity coordinates of the samples are defined as follows: L* - illumination (relative brightness); 0 – to black color and 100 – to white; a^* - tone relation between the colors red ($+a^*$) and green ($-a^*$); b^* - tone relation between the colors yellow ($+b^*$) and blue ($-b^*$); c^* - color saturation; h^* - color shade.

[1]Mariyana Sestrimska, Department Automation and Control Systems University of Food Technologies, 26 Maritza Blvd, Plovdiv, 4002 Bulgaria, e-mail: mariana_brusova@abv.bg

[2]Tanya Titova, Department Automation and Control Systems University of Food Technologies, 26 Maritza Blvd, Plovdiv, 4002 Bulgaria, e-mail: t_titova@abv.bg

[3]Veselin Nachev, Department Automation and Control Systems University of Food Technologies, 26 Maritza Blvd, Plovdiv, 4002 Bulgaria, e-mail: v_nachevbg@yahoo.com

C. VIS/NIR Spectral Analysis

Spectral characteristics of diffuse reflection in VIS/NIR range of the electromagnetic spectrum of light from 460 to 975,43nm are collected for all types of tea. Each sample is placed in a glass Petri dish with a diameter 9 cm, height 1 cm, as the liquid fills 2/3 of the volume. The light source of the spectrophotometer (FO-THLS-3100) is positioned 15 cm from Petri dish, and fiber-optic probe is placed at an angle of 45° to a height of 10 cm [3]. For each sample are measured five spectral characteristics of reflection. They were averaged to obtain one pattern.

So acquired spectral characteristics are subject to preliminary processing where the curves are smoothed by the method of moving average defined with:

$$R_{i+l/2} = \frac{1}{l+1}\sum_{k=0}^{l} R_{i+k} \ , \qquad (1)$$

where: l is width of the line filter accepting odd values; R_i are values for spectral reflection (*Reflection %*) in the range 460÷975,43nm, for different wavelengths λ.

On the basis of performed measurements is obtained so-called matrix of experiment $X_{500\times 275}$, which contains data about pH, color and spectral characteristics of diffuse reflection.

Experimental matrix is subject to PCA-analysis (Principal components analysis), in order to reduce the factor space and extracting the most valuable informative signs characterizing the samples.

D. Principal Components Analysis (PCA)

The method of principal components analysis is a mathematical procedure in which a plurality of output probability of correlated variables is transformed into a smaller number of uncorrelated variables called principal components. PCA mathematically is defined as an orthogonal linear transformation that converts the output factor space in a new coordinate system so that the greatest variation for each projection data contained in the first coordinate called the first principal component. The amount of information not described the first principal component contained in the second and so on.

The reduction of dimensionality (number of features) for the method of principal components analysis is a projection of the objects studied by multidimensional space of signs in k-dimensional space factor. This transformation is a convenient way for graphical representation and interpretation of multifactor dataset.

E. Kohonen's Self-organizing Map

Self-organizing neural networks or called Kohonen networks (Self-organizing maps, SOM) are unsupervised learning system, which are able to present the multidimensional input space in 2-D space factor called self-organizing map (SOM). Kohonen networks having two-layer structure in which the first (input) layer contains as neurons, as are sign elements and the object and the output (second) layer is connected to each neuron of the input layer. The number of neurons in the output layer is determined by the operator [4]. Fig. 1 present the basic structure of self-organizing neural network of Kohonen.

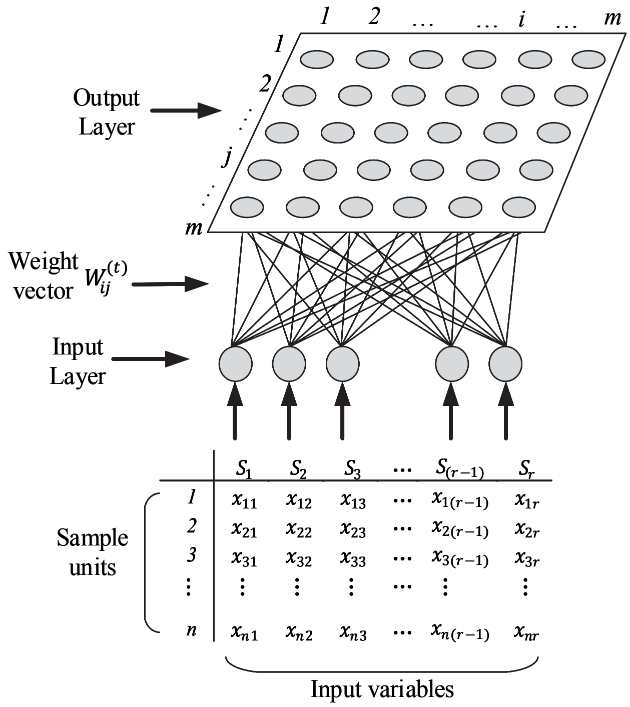

Fig. 1. Basic structure of the self-organizing map
(*n –number of analyzed samples; r - Number of input variables - factors*)

Each node j (element) of the output layer, is connected to each element i of the input layer. The weight vector $W^{(t)}$, consisting of weights $W_{ij}^{(t)}$ presents the relationship between input and output layer and adaptive changes for each iteration t of the learning process. In the initial stage of training $W^{(t)}$ is randomly and uniformly distributed in the architecture of the network. When the input of the network is submitted input vector data of each neuron of the output layer strive to achieve the greatest similarity with the elements of the input vector in the space of a predefined metric (Euclidean most often) [2]. The neurons of the output layer of the network, calculate the cumulative distance between its the weight coefficient of and the elements of the input vector, such as "winner" becomes this neuron whose weight vector represents the minimum distance to the corresponding input vector [5, 6].

III. Results and Discussion

Fig. 2 shows the average spectral characteristics of diffuse reflection for six types of tea. In addition to characteristics of reflectance of tea samples, characteristics of color (L*, a*, b*, c*, h*) and the pH were measured.

Fig. 3 is shown the location of tea samples in feature space generated by the first three principal components resulting from PCA - analysis are formed (with some overlap) six clusters corresponding to the six types of tea.

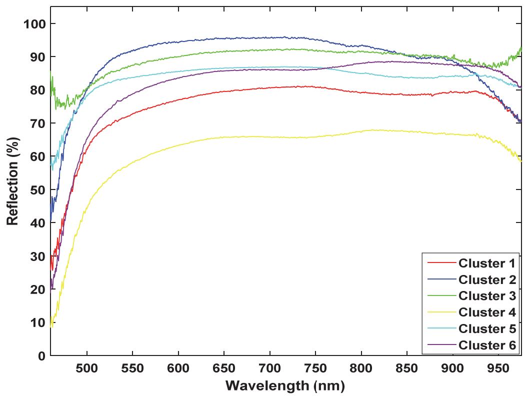

Fig. 2. Averaged spectral characteristics of six types of tea mixtures

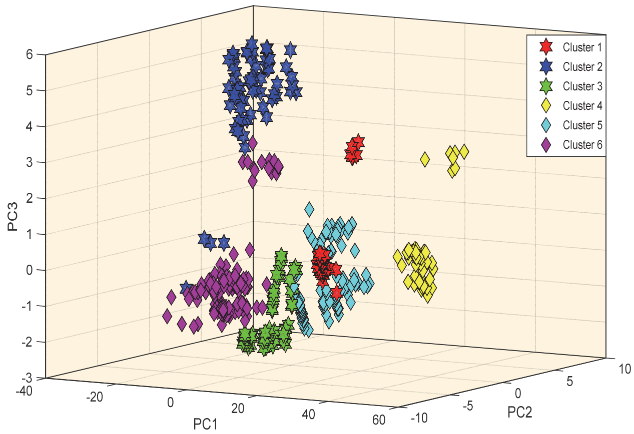

Fig. 3. Score cluster plot with first three principal components

Fig. 4 is presented the relationship between the number of principal components and the value of the eigenvectors, and the informative value of the first three principal components. The first three eigenvectors reflects the aggregate 97.87% of the variance of the output experimental data. Precisely this is the reason why they were selected as the inputs of the artificial neural network.

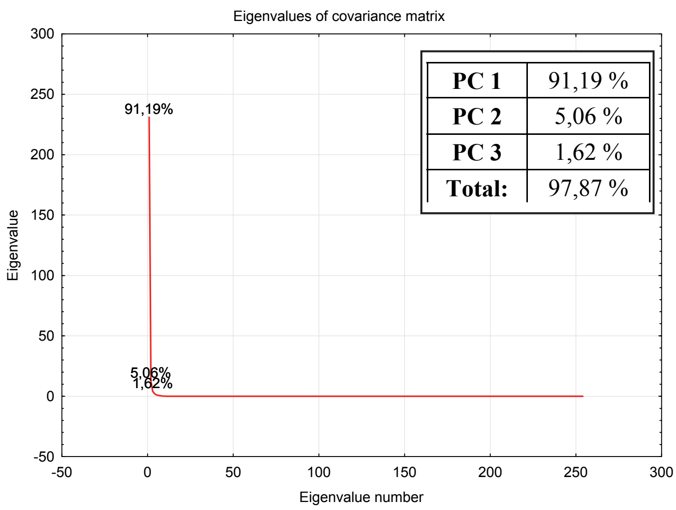

Fig. 4. Relationship between the number of principal components and the value of the eigenvectors

A. Back Propagation Artificial Neural Network (BP-ANN)

The classification of tea samples done is this paper used one-way feed-forward artificial neural network trained with the backpropagation method, using the Matlab (version R2015a).

Backpropagation algorithm aims at minimizing the mean square error of the output of the network, which is carried out using a gradient procedure [4].

In this case, a multilayer perceptron is constructed with a three-layer structure consisting of an input, an hidden layer and an output layer, and is trained with backpropagation method. As activation functions in the hidden and output network layer are used respectively *tangent-sigmoid function* and *Softmax function*. The number of input neurons of the network is three corresponding to the first three most informative principal components obtained from PCA-analysis of the output test data. In the hidden layer of neuron network has three neurons, and in the output layer 6 neurons, corresponding to six different types of tea. To each of the six groups of tea corresponds one of 6 neurons in the output layer network whose elements can take binary value 0 and 1. The unit position shows the belonging of a particular element (sample) for the respective cluster group (Table II).

TABLE II
ENCODING THE ELEMENTS IN THE OUTPUT LAYER OF THE ANN

Neurons from the output layer of the network	Binary mask
Neuron 1	1 0 0 0 0 0
Neuron 2	0 1 0 0 0 0
Neuron 3	0 0 1 0 0 0
Neuron 4	0 0 0 1 0 0
Neuron 5	0 0 0 0 1 0
Neuron 6	0 0 0 0 0 1

The percentage distribution of the experimental database is randomly as follows: 60% (325 samples) in training, 20% (100 samples) for validation, 20% (100 samples) for test. Fig. 5 illustrates the change of the mean square error training, test and validation of a neural network. Curves on validation and testing, has similar, decreasing trend. Indication of likely ineffective training would be the presence of a significant increase on test before the growth of validation curve.

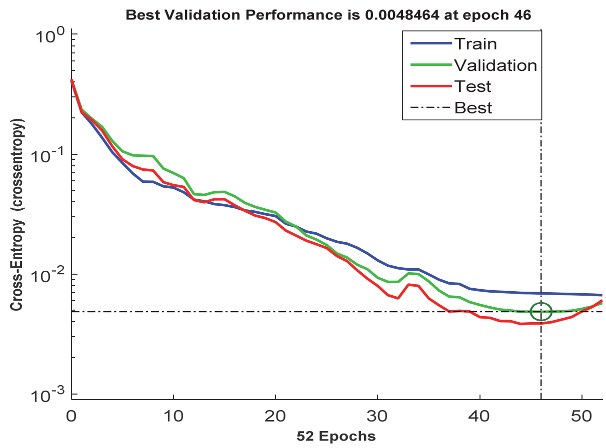

Fig. 5. Amendment of the mean square error for the learning process, test and validation of ANN

Fig. 6 presents a summary table of the errors, which shows the diagnostic accuracy of neural classifier. The table of errors is seen that a sample of cluster 1 (with a volume 50 samples) is mistakenly identified as part of the cluster 5 and two samples of cluster 5 (with a volume 120 samples) were incorrectly classified in cluster 3. The total diagnostic accuracy of the network, expressed as a percentage is 99.4%, i.e BP-ANN committed an error within 0.6%.

Output Class \ Target Class	1	2	3	4	5	6	
1	49 9.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	60 12.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	60 12.0%	0 0.0%	2 0.4%	0 0.0%	96.8% 3.2%
4	0 0.0%	0 0.0%	0 0.0%	60 12.0%	0 0.0%	0 0.0%	100% 0.0%
5	1 0.2%	0 0.0%	0 0.0%	0 0.0%	118 23.6%	0 0.0%	99.2% 0.8%
6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	150 30.0%	100% 0.0%
	98.0% 2.0%	100% 0.0%	100% 0.0%	100% 0.0%	98.3% 1.7%	100% 0.0%	99.4% 0.6%

Fig. 6. Confusion matrix

B. Kohonen's Self-organizing Map (SOM)

High-dimensional experimental feature space (500×275) is adapted to the self-organizing artificial neural network of Kohonen, where 500 are the number of samples analyzed and 275 – experimentals quality features, as colour descriptors, pH and the reflection data for different wavelengths λ. SOM has 275 input neurons.

Fig. 7 presents output layer of self-organizing neural network of Kohonen. It is a two-dimensional hexagonal card with size 10×10. The number of neurons in the network layer is determined by the operator in the Matlab programming environment. In this case, it is chosen structure 10×10 neurons in which achieved the best classification accuracy and best visual interpretation of the result achieved.

Dark hexagons represent 'winners' neurons. Their size is proportional to the number of samples strongly associated with this element. Units with low density are considered border of clusters.

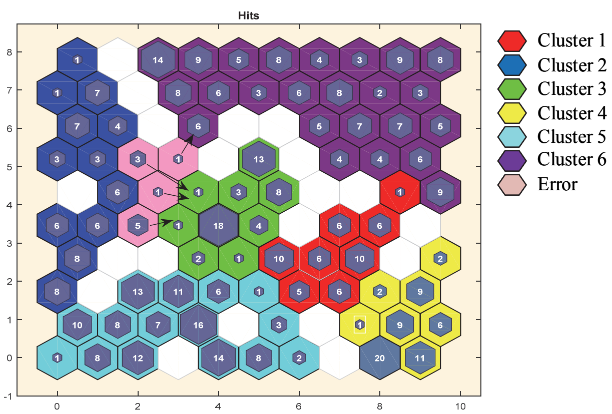

Fig. 7. Clustering through data learning by the self-organizing map

Fig. 7 shows that the network divides the tested samples of 6 cluster groups. There is overlap between the clusters 2 and 3. In this case, SOM mistakenly identified 10 tea samples. One sample is classified in cluster 3, instead of the cluster 6 and 9 samples were recognized as part of the cluster 2, instead of cluster 3. The total classification accuracy of artificial neural network is 98% (or error within 2%).

IV. Conclusion

This paper presents an approach for synthesis and training of neural classifiers for identifying different types of tea with fruit ingredients based on Backpropagation Artificial Neural Network and Kohonen's Self-organizing map. The experimental study includes collected spectral characteristics in the range 460÷975,43nm, measurement of color and pH of the tea samples. The resulting high rate of accuracy confirms the successful realization for classification task. The proposed methodology could be seen as an opportunity to identify the different types of tea with different ingredients, geographic origin and ratio of the components.

References

[1] Y. He, X. Li, and X. Deng, "Discrimination of Varieties of Tea Using Near Infrared Spectroscopy by Principal Component Analysis and BP-model", *Journal of Food Engineering*, vol. 79, pp. 1238-1242, 2009.

[2] F. Liu, X. Ye, Y. He, and L. Wang, "Application of Visible/Near Infrared Spectroscopy and Chemometric Calibrations for Variety Discrimination of Instant Milk Teas", *Journal of Food Engineering*, vol. 93, pp. 127-133, 2009.

[3] L.Gang, and H. Yang, "Discrimination of Different Brands of Nescafé Coffee Using VIS-NIR Spectroscopy and Comparative Study", *International Conference on Agricultural and Biosystems Engineering*, 2011.

[4] Ch. Damyanov, *Non-Destructive Quality Recognition in Automatic Food Sorting Systems*, UFT Academic Publishing House, Plovdiv, 2006 (in Bulgarian).

[5] M. Bieroza, A. Baker, and J. Bridgeman, "Exploratory Analysis of Excitation-Emission Matrix Fluorescence Spectra with Self-Organizing Maps as a Basis for Determination of Organic Matter Removal Efficiency at Water Treatment Works", *Journal of Geophysical Research*, 2009.

[6] T. Marique, S. Pennincx, and A. Kharoubi, "Image Segmentation and Bruise Identification on Potatoes Using a Kohonen's Self-Organizing Map", *Journal of Food Science*, 2005.