African Journal of Agricultural Research Vol. 7(21), pp. 3193-3201, 5 June, 2012
Available online at http://www.academicjournals.org/AJAR
DOI: 10.5897/AJAR11.2071
ISSN 1991-637X © 2012 Academic Journals

Full Length Research Paper

Classification of chickpea seeds using supervised and unsupervised artificial neural networks

Salah Ghamari

Department of Agricultural Machinery, Islamic Azad University, Sanandaj Branch, Sanandaj, Iran.
E-mail: sghamari2002@yahoo.com.

Accepted 4 March, 2012

As a result of varietal variability in agricultural crops, identification of seed varieties is an important problem. In this study, the ability of Artificial Neural Networks (ANN) in classification of chickpea seeds varieties was considered based on morphological properties of seeds. Experimentally, the seven morphological feature of 400 seeds (including four varieties; Kaka, Piroz, Ilc and Jam) were obtained. Using a combination of input variables, a database of 400 patterns was obtained for the development of ANN models. For comparing the supervised and unsupervised artificial neural networks in classification, the back propagation algorithm (BP) and self-organizing map (SOM) were used for classification. The results of this study showed that unsupervised artificial neural network has a better performance (with 79% accuracy and R^2 = 0.8455) in classification of chickpea varieties rather than supervised artificial neural networks (with 73% accuracy and R^2 = 0.8236).

Key words: Back propagation, classification accuracy, morphological properties, self-organizing map, variety.

INTRODUCTION

Chickpea is one of the most important pulse crops by considering either its production or consumption. The quality of chickpea seeds has distinct effect on the yield of chickpea; therefore, the proper inspection of seed quality is very important. The varietals purity is one of the factors whose inspection is more difficult and more complicated than other factors. At present, the identification of chickpea seed variety mainly depends on laboratory method and field method. These two methods have many limitations and faults, which can be noticed by the low accuracy of most of the information obtained from the field method, and the time of consumption by the laboratory method. Therefore, during the last few years, a great host of researches have been undertaken, aimed at searching for fast and reliable computational methods (such as artificial neural networks) to classify the several crops (Marini et al., 2004).

Artificial Neural Networks (ANN) models are able to learn the relationship between dependent and independent variable through the data itself without the need to develop specific functions between them (Mittal and Zhang, 2000). ANN models are often used when the relationship between parameters is unknown or very complex. It is also useful in non-linear, multivariable and non-parametric modeling. Therefore, ANN has been recently utilized in modeling the physical and mechanical properties of numerous agricultural materials. The prediction by a well-trained ANN is normally faster than the statistical models. In addition, it is possible to add or remove input and output variables in the ANN (Khazaei et al., 2005). In the past few years, there has been an increasing interest in ANN modeling in different fields of agriculture, particularly, for some areas where conventional statistical modeling failed. The applications of the ANN in agriculture include the prediction of crop yield, seeding dates, biomass production, physical and physiological damage to seeds, organic matter content in soils, aerodynamic properties of crops, estimation of sugar content in fruits, characterization of crop varieties and soil moisture estimation (Ghamari et al., 2010).

The objectives of this research were to build and evaluate the performance of supervised and unsupervised artificial neural networks for the classification of the four varieties of chickpea seeds in

Abbreviations: ANN, Artificial neural networks; **BP,** back propagation; **SOM,** self organizing map; **RMSE,** root mean square error, **MLP,** multilayer perception networks.

Table 1. Four example patterns used in artificial neural network classification.

Pattern numbers	M (g)	L (mm)	W (mm)	T (mm)	Gmd (mm)	Da (mm)	S (mm^2)	Chickpea variety
1	0.143	7.7	5.48	5.36	6.093	6.180	116.617	Kaka(1)
2	0.206	8.92	5.76	5.56	6.59	6.747	136.263	Piroz(2)
3	0.359	9.52	7.42	7.38	8.05	8.107	203.486	ILC(3)
4	0.41	9.94	7.14	8.1	8.31	8.393	217.195	Jam(4)

accordance with morphological properties.

MATERIALS AND METHODS

Data set and measurement method

In this research for classification with artificial neural network, four Iranian varieties of chickpea (Jam, ILC, Piroz and Kaka) were selected and from each variety, 100 seeds were selected randomly. Length (L), width (W), and thickness (T) of seeds are the morphological features that were measured by micrometer with 0.02 mm accuracy. The geometric mean diameter, arithmetic mean diameter and surface area of seeds were respectively calculated by using the following equation (Mohsenin, 1970):

$$D_g = \sqrt[3]{LWT} \quad (1)$$

$$D_e = \frac{(L + W + T)}{3} \quad (2)$$

$$S = \pi\left(D_g\right)^2 \quad (3)$$

Where; *Dg* = Geometric mean diameter of seed (mm); D_e = Arithmetic mean diameter of seed (mm) and *S* = surface area of seed (mm^2).

A digital scale also measured Mass (M) of seeds with 0.0001 g accuracy.

Artificial neural networks development

In this research, two types of artificial neural networks were used for classification of the chickpea varieties namely; supervised and unsupervised ANN.

Supervised artificial neural networks

The feed forward neural networks are mainly used for estimation of functions and classification of patterns. The multilayer perceptron networks (MLP) are the most commonly used feed forward ANNs. Back Propagation (BP) training algorithm is usually used for MLP network training (Menhaj, 1998). In this study, as a supervised ANN, the multilayer perceptron ANN with back propagation (BP) algorithm was selected to classify the seeds. There was a total of 400 patterns, each with eight components, seven of which were the input variables ($x_1, \ldots, x_7$), whereas, y_1 was the output variable. Four of these patterns were shown in Table 1. These patterns were randomly divided into 220, 100 and 80 patterns respectively for training, testing and verification. The training set was used to adjust the connection weights, while the testing set was used to check the performance of the model and the verification set used to determine when to stop training to avoid over-fitting (Shahin et al., 2008). To become feasible input neurons and to achieve fast convergence to minimal RMSE (root mean square error), all the datasets were normalized between 0.05 and 0.95 by using the following formula (Ghamari et al., 2010):

$$X_t = 0.05 + 0.9 * \left[\frac{(X_i - X_{min})}{(X_{max} - X_{min})}\right] \quad (4)$$

Where X_t is the normalized data for X_i, $X_{\max}$ and $X_{\min}$ are maximum and minimum of data before normalizing.

As a result of normalization, all variables acquired the same significance during the learning process. The input and target output pairs were applied to train the weights of the networks. Training process by these networks is iterative process that includes updating of weights among the different layers. During training process, the weights gradually proceed to stability. Therefore, it would be minimized error between target and predicted values. The RMSE was used to evaluate training ability of ANN models, and the performance of ANN to classification was considered using coefficient of determination (R^2) and classification accuracy (Ghamari et al., 2010):

$$RMSE = \sqrt{\frac{1}{n}\sum\left(X_m - X_p\right)^2} \quad (5)$$

$$R^2 = \frac{\left[\sum\left(X_p - \bar{X}_p\right)\left(X_m - \bar{X}_m\right)\right]^2}{\sum\left(X_p - \bar{X}_p\right)^2\left(X_m - \bar{X}_m\right)^2} \quad (6)$$

Where X_m and X_p are respectively measured and predicted data, n is the number of data and $\overline{X}$ is mean of output data.

Various ANN structures were investigated, including three and four layers with different number of neurons in each hidden layers, different values of learning coefficient and momentum, different learning coefficients and transfer functions. Once a given neural network was trained by using the appropriate training dataset, its performance was evaluated by using the testing dataset. The best ANN structure and optimum values of network parameters were obtained based on lowest error on training and test sets of data, by trial and error. The neural network professional ii/plus simulator, version 5.23 software was used in this part of study.

Unsupervised artificial neural networks

For classification of the varieties with unsupervised ANN, a self-organizing map (SOM) ANN was used. The SOM is one type of unsupervised competitive learning. Unlike supervised training, algorithms such as back propagation, unsupervised learning algorithms have no expected outputs. One advantage of the SOM is that it constantly discovers new things and changes with variable conditions and inputs. SOMs reduce dimensions by producing a map of usually one or two dimensions that plots the similarities of the data by grouping similar objects together. SOMs are particularly useful for visualization and cluster analysis in that they can be used to explore the groupings and relations within high-dimensional data by projecting the data onto a two-dimensional image that clearly indicates regions of similarity. The SOM architecture consists of two fully connected layers: an input layer and a Kohonen layer. The number of neurons in the input layer matches the number of attributes of the objects. Each neuron in the input layer has a feed-forward connection to each neuron in the Kohonen layer.

The algorithm is responsible for the formation of the SOM. First, it initializes the weights in the network by assigning them small random values. Then, the algorithm proceeds to three essential processes; competition, cooperation and adaptation (Gan et al., 2007). For considering the ability of SOM network in varieties classification, the input variables (x_1, …, x_7) of 400 patterns used in this research were applied in SOM network, and the network was asked to classify these patterns in four clusters. On the basis of lowest error on classification by trial and error, the best SOM parameters (including the number of training cycles, start and end value of learning parameter and start and end value of Sigma for the Gaussian neighborhood as % of map width) were selected. For the supervised ANN, coefficient of determination (R^2) and classification accuracy were used to consider the performance of unsupervised ANN. In this part of research, Neural Network based Clustering (Using Kohonen's Self Organizing Maps) software was used.

RESULTS AND DISCUSSION

Analysis of the statistical parameters for each variety shows that the Jam and Kaka variety respectively presented high and low mean values for all of the morphological features**.** Variation coefficients of parameters in four varieties shows that for all the morphological features there were a very small variability within the data. In all varieties, the minimum and maximum variation coefficients belong to the arithmetic mean diameter and mass of seeds respectively.

The morphological features distribution of the varieties is presented in Figure 1. This histogram shows the overlapping of each morphological property in the four varieties. The high value of this overlapping in each property is equivalent to high error in classification based on that property. Because of high overlapping in the morphological properties (Figure 1) in the four varieties, the probability of error in recognizing varieties according to each property alone is high. However, when considering these properties together, there is every probability of increasing the accuracy of varieties' recognition.

Supervised artificial neural network

Preliminary trials indicated that two hidden layer networks performed better than one hidden layer ANN in learning and predicting the correlation between input and output parameters. To determine the optimal number of neurons in hidden layers, training was used for 7-n_1-n_2-1 architectures. The number of neurons in the first hidden layer (n_1) was studied from 1 to 10 and from 0 to 10 for second hidden layer (n_2). Figure 2 shows the training performance of ANN as a function of the number of neurons in the first and second hidden layers. Results showed that among the various structures, the best training performance to classification belong to the 7-8-8-1 structure.

On the basis of the lowest error on training and test sets of data by trial and error, the best transfer function and learning rule (Figures 3 and 4) for classification were sinusoidal and delta rule. The results obtained from this research showed that the network parameters including learning coefficient and momentum values affected the ANN performances significantly but the choice of suitable learning coefficient and momentum is an important problem. The best values for learning coefficient and momentum (Figure 5) were respectively 0.1 and 0.4.

The correlation between epochs and RMSE was recorded in Figure 6. For 7-8-8-1 structure the number of epochs was increased from 5×10^2 to 3×10^4 and the amount of RMSE was calculated for training and verification datasets. As Figure 6 shows, the error on training and verification data generally decreases with increasing number of epochs, with an initial large drop in error that slows down as the network begins to learn the patterns. However, if training is allowed to continue beyond the point at which the error reaches the global minima, overtraining may arise, where memorization of the training data occurs (Khazaei et al., 2008a, b). The number of epochs was limited to 16×10^3. Table 2 shows the best structure and optimum parameters and Figure 7 shows the performance of the final supervised ANN model for the classification of the 400 chickpea seeds. The linear adjustment between the actual and predicted values gives a slope equal to 0.0.94 ($y = 0.94x + 0.32$). The resulting coefficient of determination (R^2) was 0.8236 for the regression between actual and predicted values. The classification accuracy for Kaka, Piroz, ILC and Jam varieties were respectively 76, 80, 52 and 84% respectively. The total classification accuracy for supervised ANN was 73%.

Unsupervised artificial neural network

The effect of training cycle on error of classification error

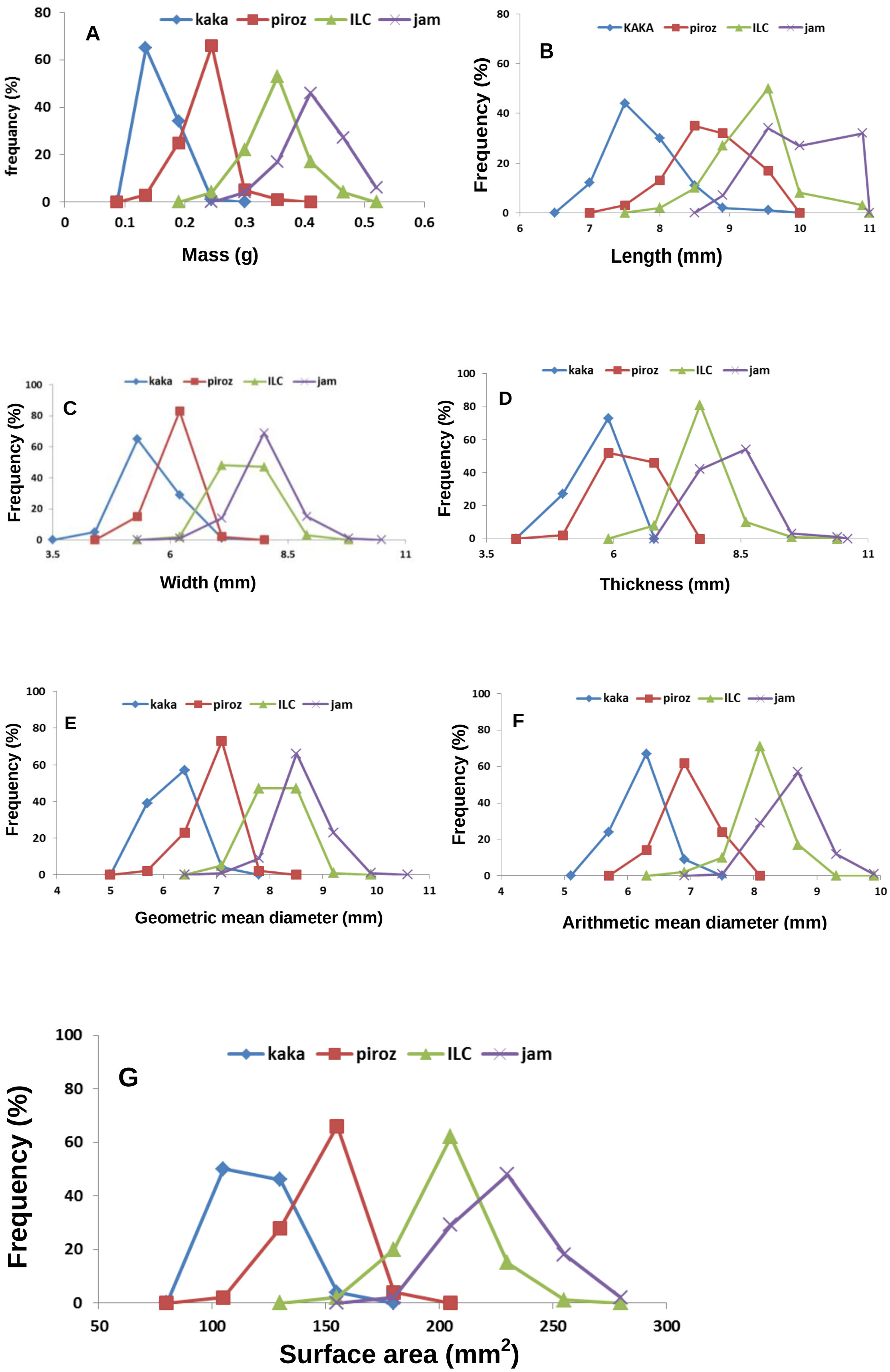

Figure 1. Frequency distribution of morphological features; A, B, C, D, E, F and G.

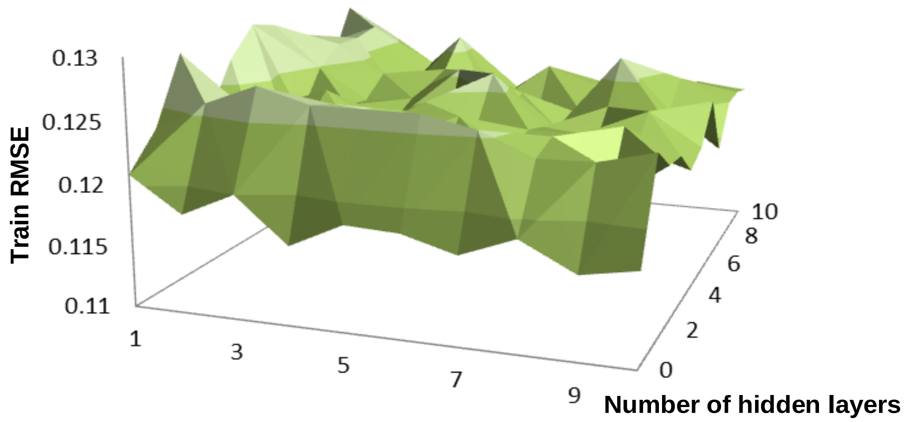

Figure 2. Training RMSE of ANN as a function of the number of neurons in the first and second hidden layers.

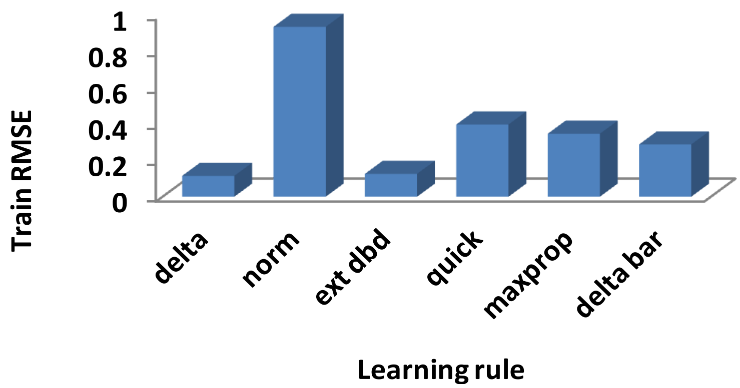

Figure 3. Comparison of training RMSE for several learning rules.

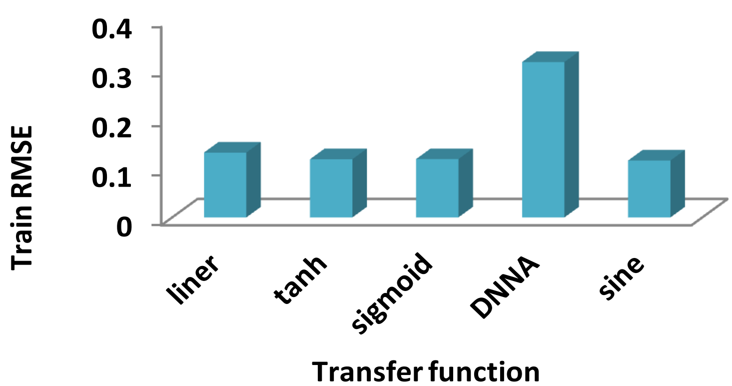

Figure 4. Comparison of training RMSE for several transfer functions.

was recorded in Figure 8. Based on the lowest error on classification by trial and error, the best number of training cycle for classification of the varieties was two cycles. To determine the optimal values of learning parameters of Kohonen's SOM, the start and end values of were respectively studied from 0.2 to 0.9 and from 0.1 to 0.8. Figure 9 shows the performance of ANN as a function of the learning parameters. Among the various values, the best performance for classification was respectively obtained when the start and end values were

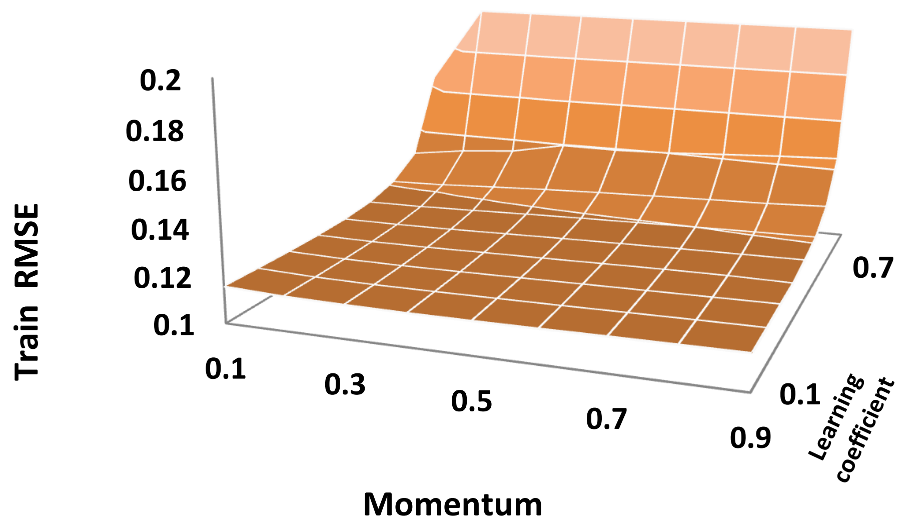

Figure 5. The effect of learning coefficient and momentum values on training performance of the ANN model.

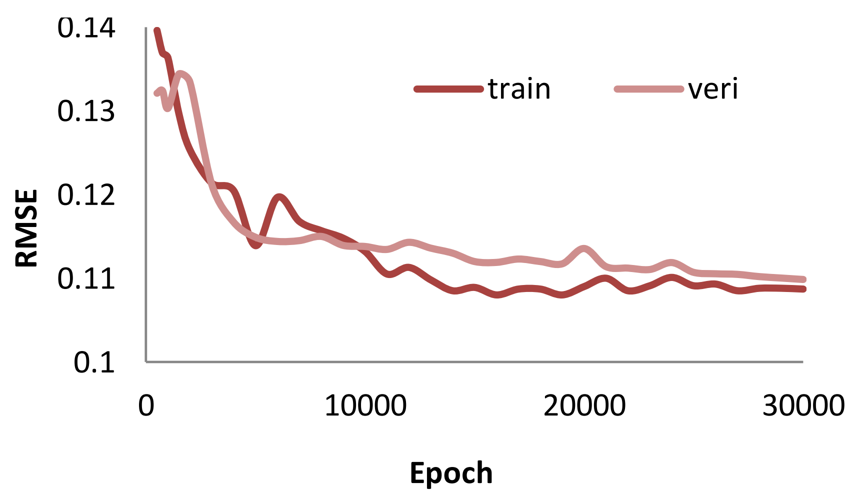

Figure 6. The training and verification RMSE as a function of the number of epochs.

Table 2. The best structure and optimum parameters used to seeds classification.

Multilayer perceptron ANN with back-propagation algorithm					
Structure	**Learning rule**	**Transfer function**	**Learning coefficient**	**Momentum**	**Epoch**
7-8-8-1	Delta rule	Sinusoidal	0.1	0.4	$16\ 10^3$

0.7 and 0.6.

The results showed that the start and end values of Sigma for the Gaussian neighborhood significantly affected the ANN performances (Figure 10), but the choice of suitable values is an important problem. In the exponential decay, desirable start and end values were respectively 50 and10%. The optimal parameters of final network are reported in Table 3. The results of the unsupervised ANN model performance for classification of the 400 chickpea seeds in four varieties was depicted in Figure 11. It shows a linear adjustment between the actual and predicted values giving a slope equal to 0.874 (y = 0.874x + 0.19). The resulting coefficient of determination (R^2) was 0.8455 for the regression between

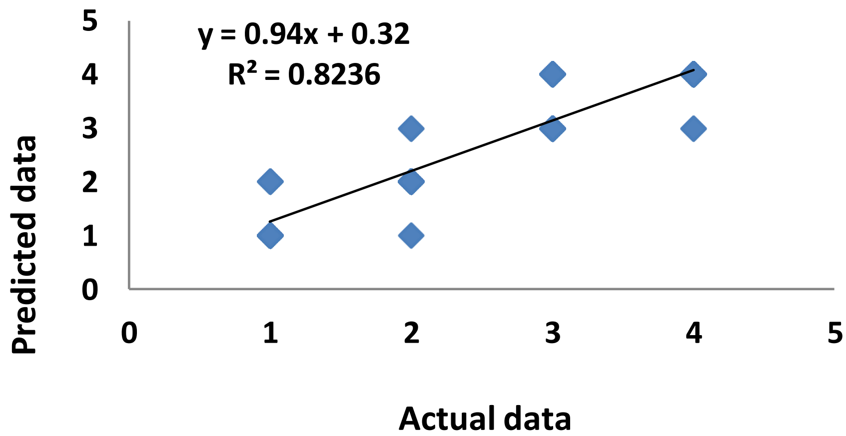

Figure 7. Correlation between the actual and predicted varieties of seeds by supervised ANN model.

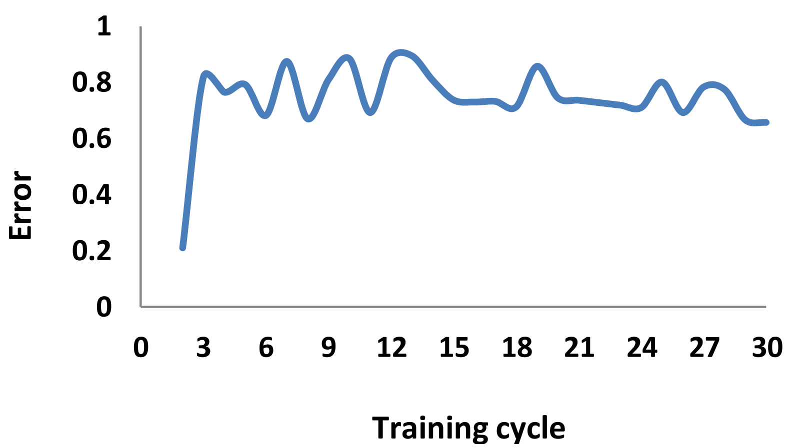

Figure 8. The effect of training cycle on error of classification error.

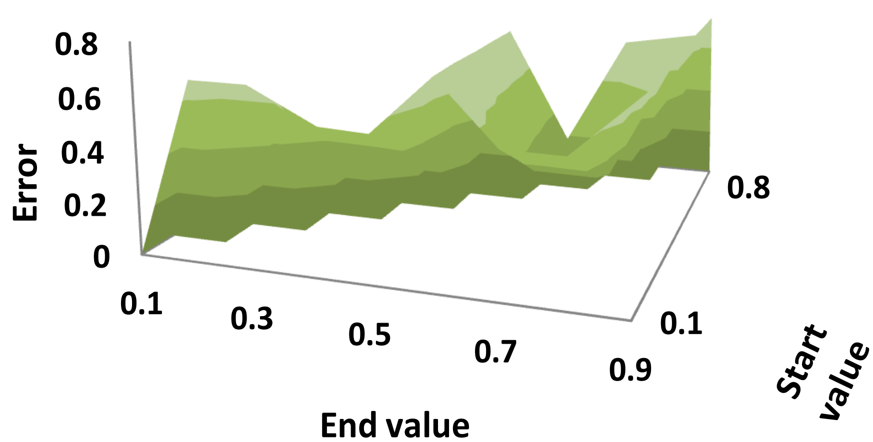

Figure 9. The effect of learning parameters on unsupervised ANN performance.

actual and predicted values. The classification error for Kaka, Piroz, Ilc and Jam variety were respectively 93, 84, 79 and 51%. The unsupervised ANN has the total classification accuracy being equal to 79%.

Table 3. The optimal parameter of final unsupervised artificial neural network.

Number of training cycle	Learning parameters		Sigma for the gaussian neighborhood		
	Start value	End value	Start value (%)	End value (%)	Decay
2	0.7	0.6	50	10	Exponential

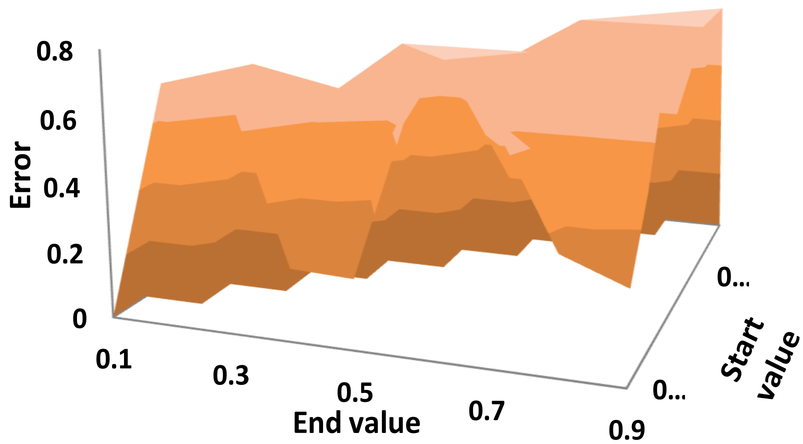

Figure 10. The classification error as a function of the start and end values of Sigma for the Gaussian neighborhood.

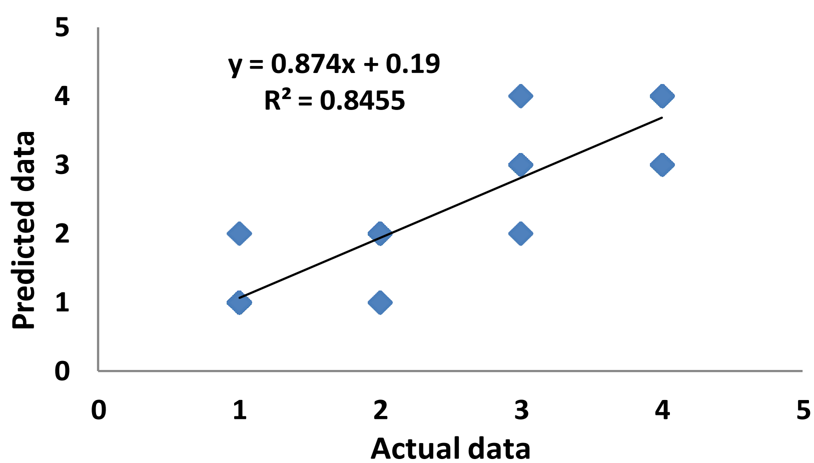

Figure 11. Correlation between the actual and predicted varieties of seeds by unsupervised ANN.

Conclusion

In this paper, the ability of a supervised (back propagation) and an unsupervised (self organizing map) artificial neural network to classify the chickpea varieties was compared. The results of this research showed that unsupervised artificial neural network has a better performance (with 79% accuracy and R^2 = 0.8455) in classification of the chickpea varieties over supervised artificial neural networks (with 73% accuracy and R^2 = 0.8236). The results also showed that, in spite of the very high overlapping of morphological properties of varieties, the accuracy of both supervised and unsupervised ANN was not low, because of the low overlapping in Kaka and Jam, but the accuracy on classification of each other was zero. Therefore, adding the other properties of seeds, such as color and texture (by using image processing), to it can improve the performance of both ANN model significantly.

REFERENCES

Gan G, Ma C, Wu J (2007), Data Clustering Theory, Algorithms, and Applications, American Statistical Association, Alexandria, Virginia.

Ghamari S, Borghei AM, Rabbani H, Khazaei J (2010), Modeling the terminal velocity of agricultural seeds with artificial neural networks. Afr. J. Agric. Res., 5(5): 389-398.

Khazaei J, Chegini GR, Kianmehr MH (2005), Modeling physical damage and percentage of threshed pods of chickpea in a finger type thresher using artificial neural networks, Journal of LucrăriŞtiinţifice, 48: 594-607.

Khazaei J, Chegini GR, Bakhshiani M (2008a), A Novel Alternative Method for Modeling the Effects of Air Temperature and Slice Thickness on Quality and Drying Kinetics of Tomato Slices, Int. Drying Technolo. J., 26: 759-775.

Khazaei J, Naghavi MR, Jahansouz MR, Salimi-Khorshidi G (2008b), Yield Estimation and Clustering of Chickpea Genotypes Using Soft Computing Techniques, Agron. J., 100(4): 1077-1087.

Marini F, Zupan J, Magr L (2004), On the use of counterpropagation artificial neural networks to characterize Italian rice varieties, Analytica Chimica Acta: pp. 231–240.

Menhaj MB (1998). Fundamental of Artificial Neural Networks, Amir Kabir University Publishin, (in Persian), Tehran, Iran.

Mittal GS, Zhang J (2000), Prediction of temperature and moisture content of frankfurters during thermal processing using neural network, Meat Sci., 55: 13–24.

Mohsenin NN (1970), Physical properties of plant and animal materials, gordon and breach, science publisher: pp. 495-513.

Shahin A, Jaksa B, Maier R (2008), State of the Art of Artificial Neural Networks in Geotechnical Engineering, EJGE.