J. Braz. Chem. Soc., Vol. 22, No. 1, 142-147, 2011.
Printed in Brazil - ©2011 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Article

Artificial Neural Networks in the Classification and Identification of Soybean Cultivars by Planting Region

Olívio F. Galão,[a] Dionísio Borsato,,[a] Jurandir P. Pinto,[a] Jesuí V. Visentainer[b] and Mercedes Concórdia Carrão-Panizzi[c]*

*[a]Departamento de Química, Universidade Estadual de Londrina, CP 6001,
86051-990 Londrina-PR, Brazil*

*[b]Universidade Estadual de Maringá, Av. Colombo 5.790, Jd. Universitário,
87020-900 Maringá-PR, Brazil*

*[c]Centro Nacional de Pesquisa de Soja (CNPSo)-Embrapa, CP 231,
86001-970 Londrina-PR, Brazil*

Vinte variedades de soja (*Glycine max*), quatorze convencionais e seis variedades transgênicas (RR) foram analisadas quanto ao teor de proteína, ácido fítico, teor de óleo, fitosteróis, cinzas, minerais e ácidos graxos que foram tabelados e apresentados à rede neural do tipo perceptron de múltiplas camadas para a classificação e identificação quanto a região de plantio e quanto a variedade convencional ou transgênica. A rede neural utilizada classificou e testou corretamente 100% das amostras cultivadas por região. Para o banco de dados contendo informações sobre sojas transgênicas e convencionais foi obtido um desempenho de 94,43% no treinamento da rede, 83,30% no teste e 100% na validação.

Twenty soybean (*Glycine max*) varieties, 14 conventional and 6 transgenic varieties were analyzed for protein content, phytic acid, oil content, phytosterols, ash, minerals and fatty acids. The data were tabled and presented to the multilayer perceptron neural network for classification and identification of their planting region and whether they were a conventional or transgenic. The neural network used correctly classified and tested 100% of the samples cultivated *per* region. For the data bank containing information on transgenic and conventional soybean, a performance of 94.43% was obtained in the training of the neural network, 83.30% in the test and 100% in the validation.

Keywords: multilayer perceptron neural networks, re-sampling, phytosterols, fatty acids

Introduction

Soybean is an important source of protein and oil and of the B complex vitamins. In this context, the development of new soybean genotypes for nutritional attributes is of great importance due to its economic value, especially for oil and protein production.[1] What most interests the market currently is the oil and protein content, but aspects such as anti-nutritional factors and others of interest to human health including oligosaccharides and phytosterols should also be considered.[2] The oil is of commercial interest because it is rich in unsaturated fatty acids that are beneficial to health and are also used for fuel production as an alternative to petroleum derivatives.[2]

Soybean proteins are present in various commercial products such as tofu, soy sauce, cholesterol-free vegetable meat and soybean milk manufacture, with a better nutritional value.[3] Furthermore, researcher's interest has been triggered because the presence of phytosterols helps to reduce cholesterol levels, phytic acid an anti-nutritional component that can complex divalent cations such as calcium and iron, making them unavailable to our organism[4] and mineral components such as zinc, copper and manganese that play a role in the enzymatic system of animals and plants.[5]

Recognizing and classifying patterns is a task naturally carried out by humans thanks to the evolution

*e-mail: dborsato@uel.br

and adaptation over thousands of years of our central nervous system, most specifically the brain cortex. However, solving classification problems by automated systems in most cases is extremely complex especially if the patterns are described by a large number of independent variables.[6,7]

The artificial neural networks (ANN) are a set of techniques based on statistical principles that have been gaining ground in pattern recognition and classification.[6,8] ANN are extremely versatile for mapping complex and nonlinear relationships among multiple input and output variables. Constructing classification rules based only on the data available, without imposing an a priori model, is another attraction of the technique.[9] The disadvantages of the methodology include the need for a large number of training data,[6] difficulty in choosing the training data and the type of neural network most suitable to the problem[10] and variable results due to the initialization and sampling.[11] However, several studies have shown promising proposals for solving or reducing the disadvantages associated to the ANN.[12-14]

The multilayer perceptron neural networks (MLP) have been successfully applied to solve several types of problems of a general nature, such as approximation, classification, categorization and prediction.[15] This flexibility has meant that this type of neural network can applied in a vast range of areas, especially for process control, weather forecasting, signal processing and image and shape analysis and recognition, voice recognition, residue treatment, ceramic engineering, fire detection, financial markets and even as support for medical diagnosis.[15-17]

They are trained under supervision by the consecrated error retropropagation algorithm that is based on the rule of learning by error correction.[7] Basically, learning by retropropagation consists of two steps through the different neural network layers: one step forwards, propagation, and one step backwards, the retropropagation. Especifically, the real response of the neural network is subtracted from a desired response (target) to produce an error signal. This error signal is then propagated backwards through the neural network, adjusting the synaptic weights so that the real response of the network moves closer to the desired response.[15,18,19]

The objective of the present study was to assess the performance of an MLP-type artificial neural network to identify and classify conventional and genetically engineered soybean seed samples, cultivated in the regions of Ponta Grossa and Londrina, Paraná State, Brazil, taking into consideration the parameters of phytosterols, minerals, phytic acid and fatty acids.

Experimental

Soybean samples

Twenty soybean varieties were studied, 6 transgenic and 14 conventional cultivars, developed by Embrapa Soybean, Londrina and recommended for planting in Central Southern Brazil in the 2008/2009 growing season. The 20 varieties were planted simultaneously in Londrina and Ponta Grossa, that have different climates, terrain and mean temperatures but received the same fertilization and irrigation treatment.

Protein content

The protein content was determined by the Kjeidhal method, according to the AOAC methodology.[20]

Phytic acid

The methodology to determine the phytic acid content was carried out according to Latta and Eskin.[21]

Fatty acids

The oil content was determined by extracting with hexane in Soxhlet extractors for 6 h and gravimetric assessment by the Clevenger method.[22] Fatty acids were determined by gas chromatography using a column with a DEGS 10% liquid phase after saponifying the oil and sterifying the fatty acids with H_2SO_4 - NH_4Cl - methanol.[23]

Minerals

The ashes were determined in triplicate, according to the T47.01 methodology.[20] The minerals contained in the ashes were analyzed according to methodology recommended[20] by that consisted of digesting the samples with 6 mol L^{-1} chloridic acid. The minerals were determined with a Perkin-Elmer Optima 3300 DV atomic emission spectrometer with plasma attached by ducts ICP OES. The work patterns were prepared by diluting stock patterns.

Phytosterol

The phytosterols were isolated using the modified procedure by Vlahakis and Hazebroek[24] and the saponified samples were washed and extracted with ethylic ether. They were determined by gas chromatography using a Shimadzu CG-17A gas chromatographer with a DB-1 column (100% dimethylpolysiloxane, polarity 5, 0.25 mm thick column, 30 m long by 0.53 mm internal diameter), initial temperature 100 °C 8 min^{-1} and 10 °C min^{-1} increase to 330 °C for 8 min, split 20, 1.2 mL min^{-1} flow, injector at 300 °C and detector at 300 °C, (ratio 1:1), sept cleaning 3 mL min^{-1}, drag gas nitrogen.

Artificial neural networks

The manual and automatic module of the artificial neural networks of the Statistics 9.0 software was used. A multilayer perceptron neural network (MLP) was used for the two cases studied containing a hidden neuron layer.[25]

The values of the oil content, proteins, phytic acid, ashes in percent, campesterol, stigmasterol, β-sitosterol in mg 100 g^{-1}, calcium, magnesium, phosphorus, potassium and sulfur in g kg^{-1} samples, manganese, zinc, copper, iron and borum in mg kg^{-1} sample and C14:0, C16:0, C16:1n9, C17:0, C18:0, C18:1N9, C18:1N7, C18:2N6, C18:3n3, C20:0 and C24:0 in mg kg^{-1} were tabled and presented to the neural network in the order presented.

Computer processing

All the results of experiments were processed in an Intel Pentium Core 2 Duo with 1.83 GHz and 2.0 Gb RAM memory.

Results and Discussion

A multilayer perceptron neural network was used, consisting of an input layer containing a neuron for each one of the 28 input variables and a single intermediary layer responsible for separating the patterns by decision frontiers.

For the first case studied, classification by region, the neural network used was formed by three hidden neurons, with a learning rate of 0.04 at 100 training epochs. The learning process was maintained until the synaptic weights and the bias levels stabilized and the mean quadratic error converged to a minimum value.[26]

The constant moment was little influenced by the choice of the activation function or normalization methodology because its value remained around 0.3 for the neural network studied.

The sequential training mode was used for the neural network because the weights are updated when every new example is presented.[27] Before being fed to the neural network, all the entry variables were transformed to a scale from zero to 1 (minimax) with the function of logistics-type activation.[25] The exit layer contained two neurons, the first for the Londrina samples and the second for the samples from Ponta Grossa. The retropropagation algorithm[8] was used for the neural network constructed and the training sample sequence was randomized for each epoch. Weight correction was based on the sum of the quadratic error of the neural network and was carried out after presenting each training example.[6]

The samples were divided into three parts and the first consisted of the neural network training set and was formed with 70% of the samples. The second, called test, consisted of 15% and the third, also consisting of 15% of the samples, was called validation and all were chosen randomly. The objective of the last set was to verify the capacity for generalization of the trained neural network, because to classify, the neural network learns a rule using the training examples.

An order of importance for the entry variables was stipulated from the trained neural network.[27] The C14:0 content was identified as the most important for the neural network, followed by the contents of calcium, zinc, C24:0, campestrol and manganese as the six most important, that is, the ones that most influenced classification by planting region. The contents of C18:2n6, C18:1n7, stigmasterol, phytic acid, oil and β-phytosterol contents, because of the small variability in the data, were the least important.

The surface response contours generated using the C14:0 and calcium content according to the producing region are shown in Figure 1. It shows the influence exerted by these independent variables and that an increase in the C14:0 and calcium contents indicated that the soybean, regardless of whether it was a conventional or genetically engineered variety, was produced in the Londrina region.

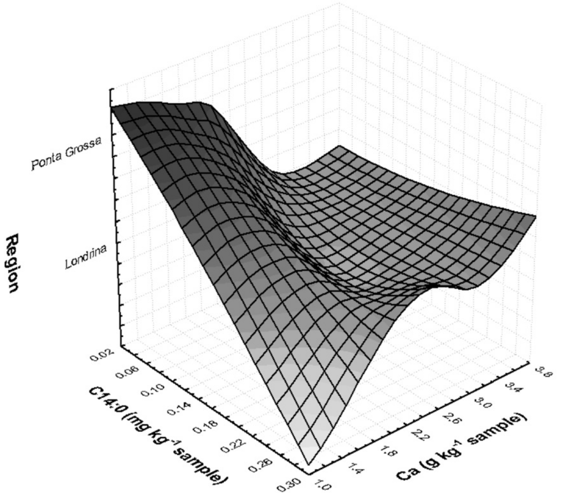

Figure 1. Contour region showing the influence of the C14:0 and calcium contents on sample classification by planting region.

Indeed, the mean contents of C14:0, calcium and β-sitosterol observed experimentally in the samples cultivated in Ponta Grossa corresponded to 42%, 60% and 92%, respectively, of the content observed in the samples cultivated in Londrina, without taking the varieties into consideration. However, the neural network did not consider the analysis of the mean value of any parameter. It was trained, using a data set, to identify the characteristics and with this classify the samples.[25]

In spite of the small number of neurons in the hidden layer, all the soybean samples were classified correctly by the neural network, according to the producing region, showing 100% accuracy in training, in the test and validation.

In the second analysis, the neural network was also fed with the 28 established parameters and for this case the automatic module was used also using the multilayer perceptron-type neural network. Here the samples were also divided into three parts and the first consisted of the training set of the neural network that was formed with 70% of the samples of the four types of soybean used. The second, called test, consisted of 15% and the third, called validation, of 15% of the samples and the three sets were chosen randomly.

The bootstrap resampling technique was used generating a greater number of random data with the same means and standard deviation as the original.[27]

Figure 2 shows the architecture of the neural network used with all the nodes of the input layer connected to the neurons of the hidden layer and these connected with the output. It also shows the first output neuron, the most activated, indicating the data of the samples cultivated in Londrina. It presented connection among the neurons of the hidden layer and showed that the most active were the fourth and fifth followed by the four last neurons. There was an associated weight for each connection among the neurons.

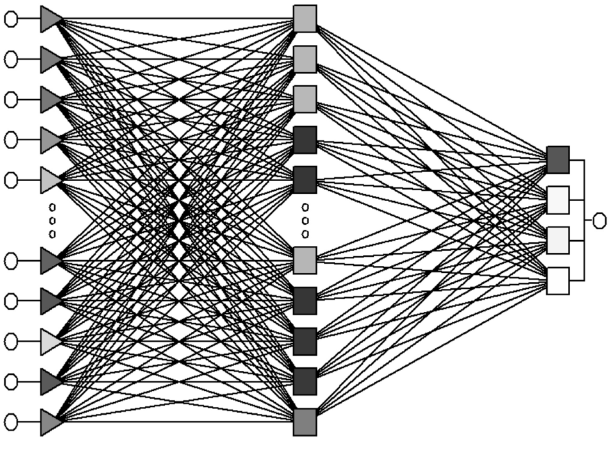

Figure 2. Graphic representation of the multilayer perceptron neural network with a hidden layer (MLP28:14:4).

The output of the first neural network was four conventional and genetically engineered soybean cultivated in Londrina, the third and fourth for conventional and transgenic soybean cultivated in Ponta Grossa. The MLP limits for the number of neurons was a minimum of 7 and maximum of 21. Twenty neural networks were trained and the best, which presented greatest accuracy in training (94.43%) and the test (83.30%) was the one with 14 neurons in the hidden layer, that is, MLP28:14:4. The neural network used was acceptable indicating that it could preserve the knowledge acquired during training and could make an impartial estimates of the neural network performance because the validation reached the index of 100%.[31]

The trained neural network also stipulated an order of importance for the input variables. The contents of zinc, manganese, C14:0, protein, calcium and C24:0 were identified as the most important in the sample classification while the contents of sulfur, oil, C18:2n6, C16:1n9 and C18:0 were, in the order presented, those of least importance.

Figure 3 shows the mean contents of manganese and zinc, variables considered more important by the neural network, in the conventional and transgenic soybean in the region of Londrina and Ponta Grossa. It shows that the mean manganese and zinc contents were greater in the samples cultivated in the Londrina region.

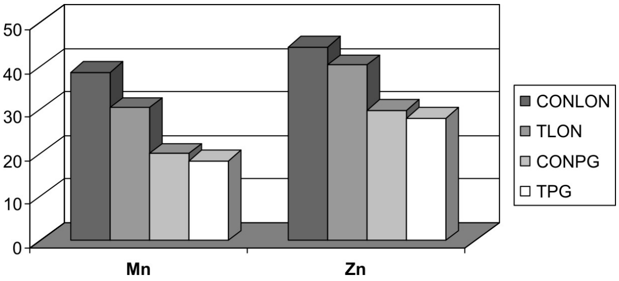

Figure 3. Mean value of the manganese and zinc contents of the samples of conventional and transgenic soybean seeds cultivated in Londrina and Ponta Grossa (CONLON: Londrina conventional, TLON: Londrina transgenic, CONPG: Ponta Grossa conventional and TPG: Ponta Grossa transgenic).

Figure 4 shows the surface response contours generated using the manganese and calcium contents, according to the producing region. It shows that lower calcium and manganese contents indicated that the soybean was planted in the Ponta Grossa region. Intermediate values of the calcium content and low manganese values indicated transgenic soybean cultivated in the Londrina region. The greatest calcium and manganese contents were observed in the samples cultivated in Londrina and the mean calcium value in the transgenic soybean from Londrina was similar to that of conventional soybean. However, it was, respectively, 85 and 61.92% greater than in the transgenic and conventional soybean cultivated in Ponta Grossa. A similar performance was observed when the mean zinc and manganese contents were compared that were significantly greater than in the soybean planted in Londrina.

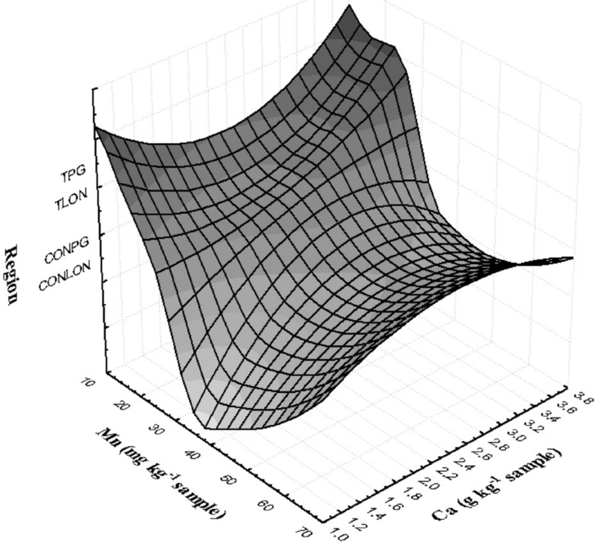

Figure 4. Contour region showing the influence of the manganese and calcium contents in the classification of samples by variety and region. (CONLON: Londrina conventional, TLON: Londrina transgenic, CONPG: Ponta Grossa conventional and TPG: Ponta Grossa transgenic).

Phytosterol are isoprenoids widely distributed in plants and the content and composition of the phytosterol that are enzymatically synthesized from Acetyl-CoA are different among the plant species. They appear in oil-rich seeds such as sesame, soybean, sunflower and rape seed.[31,32] It has recently been demonstrated that adding soybean phytosteroid to some foodstuffs reduces cholesterol absorption thus improving human health.[31]

The phytosterol most studied are β-sitosterol, stigmasterol and campesterol that have a structure similar to cholesterol.[32] The values of the mean contents obtained, in the soybean varieties studied, regardless of the planting region, were 55.52, 57.87 and 136.54 mg *per* 100 g for campesterol, stigmasterol and sistosterol, respectively. These values were greater than those obtained by Yamaya *et al.*.[32] Furthermore, some authors have shown that the protein content is related to the oil content and this to the phytosterol content in soybean samples.[24,31]

Figure 5 shows the oil and protein contents of the different soybean varieties cultivated, with the β-sitosterol content. There was an inverse ratio between the oil and protein contents, that is, the quantity of β-sitosterol increased with the increase in the oil content and decease in protein content. However, an increase in β-sitosterol content was also observed when the protein content increased and the oil content decreased.

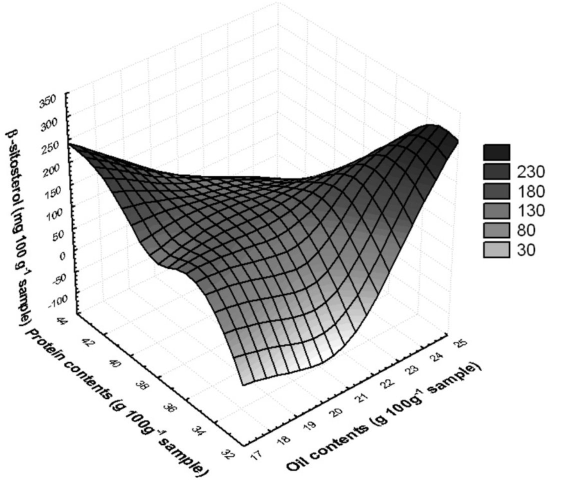

Figure 5. Contour region showing the influence of the protein and oil contents on the β-sitosterol content present in the soybean varieties studied.

The performance of the different varieties may be analyzed by the contour figures generated using artificial neural networks. The figures showed a tendency to agreement with the varieties, parameter and regions studied.

Conclusions

Soybean is a functional food because of its high protein content, unsaturated fats and other substances of interest for human health. The contents of these substances vary from region to region depending on the climate and the terrain where the soybean is cultivated.

The use of multilayer perceptron artificial neural networks was shown to be a useful tool for identifying soybean varieties cultivated in two regions of Paraná State because 100% of the samples analyzed could be recognized when the two regions were compared. For the data bank containing information on transgenic and conventional soybean a 94.43% performance was obtained in the training of the neural network, 83.30% in the test and 100% in the validation, showing the efficaciousness of the proposed classifier.

Acknowledgments

The author wishes to thank the Fundação Araucária, UEM, UEL and Embrapa-CNPSo for their financial support.

References

1. Kinney, A. J.; *J. Food Lipids* **1996**, *3*, 273.
2. Ferrari, R. A.; Oliveira, V. S.; Scabio, A.; *Quim. Nova* **2005**, *28*, 19.
3. Sridhara, S; Thimmegowda, S; Chalapatri M. V.; *Crop Res.* **1997**, *13*, 259.
4. Martinez, B. D.; Gomez, I.; Leon, M. V. R; Rincón-Leon, F.; *Arch. Latinoam. Nutr.* **2002**, *52*, 219.
5. Lee, J.; *Concise Inorganic Chemistry*; Chapman & Hall: Londres, 1966.
6. Bishop, C. M.; *Neural Networks for Pattern Recognition*, 1st ed.; University Oxford: Oxford, 1995.
7. Haykin, S.; *Redes Neurais: Princípios e Práticas*.; 1ª ed.; Bookman: Porto Alegre, Brasil, 2001.
8. Warner, B.; Misra, M.; *The Am. Stat.* **1996***, 50*, 284.
9. Sablani, S. S.; *Compr. Rev. Food Sci. Food Saf.* **2008**, *7*, 130.
10. Curry, B.; Morgan, P. H.; *Eur. J. Op. Res*. **2006**, *170*, 567.
11. Bodt, E.; Cottrell, M.; Verleysen, M.; *Neural Networks* **2002**, *15*, 967.
12. Looney, C. G.; *IEEE Trans. Know. Data Eng.* **1996**, *8*, 211.
13. Ludermir, T. B.; Yamazaki, A.; Zanchetin, C.; *IEEE Trans. Neural Networks* **2006**, *17*, 1452.
14. Windeatt, T.; *IEEE Trans. Neural Networks* **2006**, *17*, 1194.
15. Braga, A. P.; Carvalho, A. C. P. L. F.; Ludermir, T. B.; *Redes Neurais Artificiais: Teoria e Aplicações*, 1ª ed.; LTC: Rio de Janeiro, Brasil, 2000.
16. Mukesh, D.; *J. Chem. Educ*. **1996**, *73*, 431.
17. Mandal, S.; Prabaharam, N.; *Ocean Eng.* **2006**, *33*,1401.
18. Hirose, Y.; Yamashita, K.; Hijiya, S.; *Neural Networks* **1991**, *4*, 61.
19. Hilera González, J. R.; Martinez Hernando, V. J.; *Redes Neuronales Artificiales: Fundamentos Modelos y Aplicaciones*, 1ª ed.; Editorial Alfaomega: Madrid, 2000.
20. AOAC-*Association of Official Analytical Chemists*, Official methods of analysis of the Association of Official Analytical Chemists; 14th ed.; Washington: DC, 1984.
21. Latta, M.; Eskin, M.; *J. Agric. Food Chem.* **1980**, *28*, 1313.
22. AOAC-*Association of Official Analytical Chemists*; Official methods of analysis of the Association of Official Analytical Chemists; Washington: DC, 1990.
23. Hartman, L.; Lago, R. C. A.; *Lab. Pract.* 1973, *22*, 475.
24. Vlahakis, C.; Hazebroek, J.; *J. Am. Oil Chem. Soc.* **2000**, *77*, 49.
25. *Statistica V 9.1 For Windows*; Statsoft Inc. Software, Tulsa, 2009.
26. Azevedo, F. M.; Brasil, L. M.; Oliveira, R. C. L.; *As Redes Neurais com Aplicações em Controle e em Sistemas Especialistas***,** 1ª ed.; Bookstores: Florianópolis, Brasil, 2000.
27. Borsato, D.; Moreira, I.; Nóbrega, M. M.; Moreira, M. B.; Dias, G. H., Silva, R. S. S. F., Bona, E.; *Quim. Nova* **2009**, 9, 2328.
28. Hunter, A.; Kennedy, L.; Henry, J.; Fergunson, I.; *Comput. Methods Prog. Biomed.* **2000**, *62*, 11.
29. Tudu, B.; Jana, A.; Metla, A.; Gosh, D.; Bhattacharyya, N.; Bandyopadhyay, R.; *Sens. Actuators, B* **2009**, *138*, 90.
30. Yamaya, A.; Endo, Y.; Fujimoto, K.; Kitamura, K.; *Food Chem.* **2007**, *102*, 1071.
31. Law, M. R.; *Br. Med. J.* **2000**, *320*, 861.

Submitted: May 14, 2010
Published online: August 31, 2010