USE OF COMPUTATIONAL INTELLIGENCE IN THE GENETIC DIVERGENCE OF COLORED COTTON PLANTS

Daniel Bonifácio Oliveira CARDOSO[1] iD, Luiza Amaral MEDEIROS[1] iD, Gabriela de Oliveira CARVALHO[2] iD, Izabela Motta PIMENTEL[2] iD, Gabriella Xavier ROJAS[2] iD, Lara Araújo SOUZA[2] iD, Gabriel Medeiros SOUZA[2] iD, Larissa Barbosa de SOUSA[2] iD

[1] Postgraduate Program in Agronomy, Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil.
[2] Institute of Science Agrariam, Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil.

Corresponding author:
Daniel Bonifácio Oliveira Cardoso
Email: danieludia13@hotmail.com

Abstract

The objective of this work was to analyze the genetic diversity using conventional methods and artificial neural networks among 12 colored fiber cotton genotypes, using technological characteristics of the fiber and productivity in terms of cottonseed and cotton fiber yield. The experiment was conducted in an experimental area located at Fazenda Capim Branco, belonging to the Federal University of Uberlândia, in the city of Uberlândia, Minas Gerais. Twelve genotypes of colored fiber cotton were evaluated, 10 from the Cotton Genetic Improvement Program (PROMALG): UFUJP - 01, UFUJP - 02, UFUJP - 05, UFUJP - 08, UFUJP - 09, UFUJP - 10, UFUJP - 11, UFUJP - 13, UFUJP - 16, UFUJP - 17 and two commercial cultivars: BRS Rubi (RC) and BRS Topázio (TC). The experimental design used was complete randomized block (CRB) with three replications. The following evaluations were carried out at full maturation: yield of cottonseed (kg ha-1) and the technological characteristics, which include, fiber length, micronaire, maturation, length uniformity, short fiber index, elongation and strength, using the HVI (High volume instrument) device. Genetic dissimilarity was measured using the generalized Mahalanobis distance and after obtaining the dissimilarity matrix, the genotypes were grouped using a hierarchical clustering method (UPGMA). A discriminant analysis and the Kohonen Self-Organizing Map (SOM) by Artificial Neural Networks (ANN's) were performed through computational intelligence. SOM was able to detect differences and organize the similarities between accesses in a more coherent way, forming a larger number of groups, when compared to the method that uses the Mahalanobis matrix. It was also more accurate than the discriminant analysis, since it made it possible to differentiate groups more coherently when comparing their phenotypic behavior. The methods that use computational intelligence proved to be more efficient in detecting similarity, with Kohonen's Self-Organizing Map being the most adequate to classify and group cotton genotypes.

Keywords: *Gossypium hirsutum*. Kohonen Self-Organizing Maps. Neural Networks.

1. Introduction

Cotton is grown in more than 72 countries on five continents with more than 90% of world's production is of the *Gossypium hirsutum* species, with a large part consisting of white fiber (Borém and Freire 2014). Cotton is considered the most important natural textile fiber in the world, as it is used to dress almost half of the global population (Cardoso et al. 2019).

In Brazil, it is an important commodities in the agriculture. It is the fourth producer in the world and the second in export volume, with emphasis on Mato Grosso, Bahia and Minas Gerais as the largest producers in the country (ABRAPA 2020).

However, cotton plants produce colored fibered cotton naturally, which has a small niche market. This naturally colored cotton is important since the fiber does not need to be dyed, eliminating the use of water and reducing production costs (Dutt et al. 2008). However, these fibers are of low quality when compared to white fibers, and therefore research into genetically improving the plants is required (Cardoso 2019).

In this sense, one of the pillars of plant breeding is genetic diversity, as it makes it possible to identify superior hybrid combinations with greater heterotic effect and greater heterozygosity, in order to find genotypes with characteristics of interest (Cruz et al. 2014).

The diversity among parent plants is usually measured using techniques that use biometric models, by cluster analysis methods, main components or canonical variables. On the other hand, there is computational intelligence, which uses models that simulate the human brain, where learning is done through mistakes, successes and experiences (Cruz and Nascimento 2018).

Computational intelligence is an alternative to conventional analysis. It has the advantage of being non-parametric, analyzing the data even if they are unbalanced, have experimental errors and contain flaws in the assumptions (Cruz and Nascimento 2018). Among the techniques used in plant breeding, artificial neural networks, Fuzzy logic and evolutionary computing stand out.

Artificial neural networks (ANN's) simulate human behavior, with neurons and synapses transmitting information (estimating weights between them), making mistakes and getting it right, learning from experience and making decisions. In plant breeding, it is used to classify and group genotypes, in genetic diversity, prediction of genetic value, adaptability and stability, among other things (Haykin 2008; Nascimento et al. 2013; Oliveira et al. 2013; Bhering et al. 2015; Cardoso et al. 2019).

In cotton culture, ARR has been shown to be efficient. Hu et al. (2019), analyzing the cotton yarn quality prediction model based on the artificial recurrent neural network, found that the experimental results show better accuracy. Cardoso et al. (2019), found greater efficiency of ARR for studies of adaptability and stability in cotton, when compared to conventional statistics.

One class of ANN's is Kohonen's Self-Organizing Maps (SOM) that recognizes patterns, clusters and data organization (Cruz and Nascimento 2018) detecting the dissimilarity between genotypes through competitive learning, determining weights for the winning neuron and a radius establishes its neighborhood, with neurons being classified as individuals.

SOM is used in several areas of scientific knowledge. Rodrigo et al. (2012), using SOM observed consistency when checking the gait of individuals with Parkinson's disease. Silva (2018) used SOM to estimate genetic divergence in corn, finding divergence between the use of ANN in relation to multivariate methods.

Based on the above, the objective of this work was to analyze genetic diversity through conventional methods and artificial neural networks among 12 colored fiber cotton genotypes, using the technological characteristics of the fiber, yield and productivity of cotton and cottonseed.

2. Material and Methods

The experiment was conducted in an experimental area located at Fazenda Capim Branco (18º52'S; 48º20'W and 805m altitude), belonging to the Federal University of Uberlândia, in the municipality of Uberlândia, Minas Gerais in the 2013/14, 2014/15, 2015/16, 2016/17 and 2017/18 seasons. The city has an average air temperature of 22.4ºC, an average relative humidity of 70% and an average annual rainfall of 1,584 mm per year. The area where the experiment was carried out is a dystrophic Dark Red Latosol, with a clay texture.

Twelve genotypes of colored fiber cotton were evaluated, 10 from the Cotton Genetic Improvement Program (PROMALG): UFUJP - 01, UFUJP - 02, UFUJP - 05, UFUJP - 08, UFUJP - 09, UFUJP - 10, UFUJP - 11, UFUJP - 13, UFUJP - 16, UFUJP - 17 and two commercial cultivars: BRS Rubi (RC) and BRS Topázio (TC).

The experimental design used was complete randomized blocks (CRB) with three replications. The experimental plot consisted of four lines of five meters, spaced one meter apart, with the useful area being composed of the two central lines neglecting 0.5 m from each end of the line.

At full maturity, the weight of cottonseed (kg ha^{-1}) and fiber yield were evaluated. The technological characteristics of the fiber were analyzed in the fiber quality analysis laboratory of the Minas Gerais Association of Cotton Producers (AMIPA). These technological characteristics were fiber length, micronaire, maturation, length uniformity, short fiber index, elongation and strength, with the aid of the HVI (*High volume instrument*) device.

The micronaire index (MIC),fiber maturity (MAT the strength (STR) gf tex^{-1}; the fiber length (UHML)in mm; length uniformity (UI) in %; the short fiber index (IFC)in mm; fiber elongation (ALG)in mm.

The data were submitted to univariate and multivariate analysis of variances and, from this, the means were obtained to perform the analyzes. The genetic dissimilarity between the pairs of genotypes using the Generalized Mahalanobis Distance (D$^2_{ii'}$) were estimated as below:

$$D^2_{ii'} = \delta' \Psi^{-1} \delta$$

In which:

D$^2_{ii'}$: generalized Mahalanobis distance between the genotypes i and i';

Ψ: matrix of variances and residual covariance;

δ′: [d$_1$ d$_2$ ...   d$_v$] where d$_j$ = Y$_{ij}$ − Y$_{i'j}$;

Y$_{ij}$: mean of the i-th genotype in relation to the j-th variable.

After obtaining the dissimilarity matrix, the genotypes were grouped using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), generating a dendrogram of greatest similarity in which the distance between the genotype and the group formed by individuals i and j is given by:

$$d_{(ij)k} = \frac{d_{ik} + d_{jk}}{2}$$

Through computational intelligence, discriminant analysis and Kohonen's Self-Organizing Map (SOM) were performed using Artificial Neural Networks (ANN's). The architecture of SOM was of the *feedforward* type with an input layer and an output layer, called topological map which is divided into three stages (Cruz and Nascimento 2018):

1st Stage: Definition of the topological map and establishment of random weights. The following parameters were used for the formation of SOM: 3 neurons in two dimensions (Figure 1) (three rows and 3 columns), 2000 times and radius neighborhood pattern = 2 and the *dist* activation function (Euclidean distance), and topology of the hexagons. Afterwards, the synaptic weights and an input vector Xi will start.

2nd Stage: Given the input values, the measurement of the distance in competition was calculated, and the winning neuron was established as the one with the shortest distance between it and the input data, and the neighboring neurons had their weights adjusted in relation to the input, to determine the neighborhood for the rate of learning (η), and was determined by the following expression:

$$i.\, w^{i+1}(\text{winner}) = W^i + (\text{winner})^{+\eta(Xi-W^1)}(\text{winner})$$
$$ii.\, w^{i+1}(\text{neighborhood}) = W^i + (\text{neighborhood})^{+f(x)\eta(Xi-W^1)}(\text{neighborhood})$$

η= measurement of the learning rate; w = weight of neurons; xi= input vector; f(x)= half of the learning rate.

3rd Stage: Each input participates in the competition, ending one time and stage 2 is resumed when there are no major changes between the weights of input and actuals.
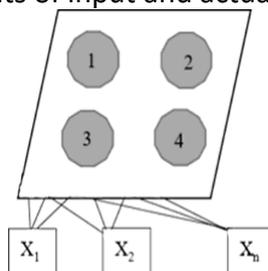
Figure 1. Architecture and topology of a SOM neural network in two dimensions (Adapted from Cruz and Nascimento 2018).

The discriminant analysis was performed by means of ANN's using a neural network of the *Multilayer Perceptron* (MLP) type formed by two layers containing between two and five neurons in each layer, using the logarithmic activation function. The training algorithm chosen was Trainlm (*Levenberg-Marquardt backpropagation*). The training cycle was set at 5000 times and an error rate of 0.01. The network had 1000 observations and separated 80% of the data for training and 20% for validation (Figure 2).
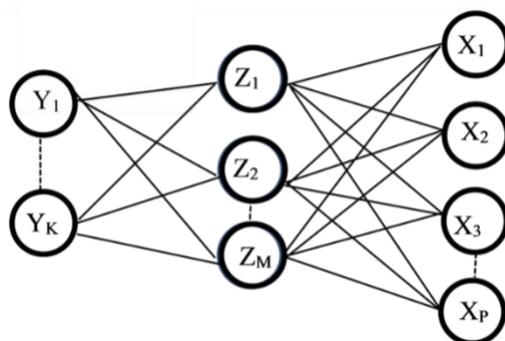
Figure 2. Scheme of the single hidden layer of the neural network (Adapted from Nascimento et al. 2013).

The analyzes were performed using the statistical program (GENES), integrated with the R and Matlab software (Cruz 2016).

## 3. Results and Discussion

The means of the characteristics demonstrate the formation of groups for all characteristics, therefore, there is variability between the evaluated genotypes. In general, commercial genotypes had the best averages, with the exception of elongation, which shows that the variability needs to be explored between these genotypes (Table 1), due to the responsiveness of PROMALG genotypes to the environment.

Table 1. Average of seven characteristics of 12 colored fiber genotypes grown in Uberlândia-MG.

| Genotypes	UHML	UI	IFC	STR	ALG	CSY	FP%
UFUJP-01	24.29b	79.06a	14.04b	22.16b	8.95a	1842.88c	29.74b
UFUJP-02	24.42b	78.36b	13.87b	22.58b	8.33a	2342.64b	28.66b
UFUJP-05	23.83b	77.76b	15.78b	21.33b	8.99a	1845.63c	29.54b
UFUJP-08	23.73b	78.14b	15.52b	21.37b	8.62a	2032.60c	29.84b
UFUJP-09	23.51b	77.19b	15.90b	21.07b	8.53a	2283.07b	29.13b
UFUJP-10	24.34b	77.65b	14.82b	22.05b	8.53a	2129.20b	28.72b
UFUJP-11	23.60b	78.38b	14.63b	21.16b	9.07a	2203.76b	29.57b
UFUJP-13	23.74b	77.83b	14.58b	22.25b	8.34a	1829.41c	30.67b
UFUJP-16	23.42b	77.54b	15.08b	21.59b	8.71a	2364.51b	32.13b
UFUJP-17	24.36b	77.21b	15.05b	21.65b	8.72a	2494.81b	30.28b
BRS RUBI (C)	26.07a	79.60a	13.05a	24.99a	7.84a	2908.46a	33.43a
BRSTOPÁZIO (C)	27.17a	79.21a	11.32a	25.45a	7.66a	2829.15a	34.39a

Averages followed by the same capital letter (horizontally) or the same lowercase letter (vertically) do not differ by Scott-Knott's test at 5% probability. UHML: fiber length (mm); UI: length uniformity (%); IFC: Short Fiber Index; STR: resistance (gf/tex); ALG: Elongation; CSY: cottonseed yield (kg/ha); FP: fiber percentage (%). (C): Commercial Variety.

For the dendrogram, the cut was made considering the abrupt change in level (Miranda 2019). With this it is possible to observe the formation of four distinct groups (Figure 3). There is also a co-phenetic correlation of 0.95, which indicates a good graphic representation of the dendrogram, based on data from the dissimilarity matrix (Nardino et al. 2017).
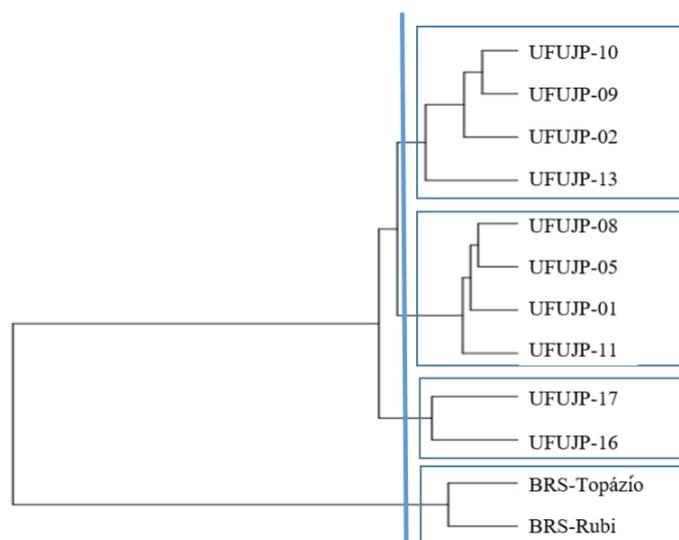
Figure 3. Dendrogram of genetic divergence between 12 cotton genotypes, obtained by the unweighted pair group method with arithmetic mean "UPGMA", based on the generalized Mahalanobis distance (D^2). Co-phenetic correlation coefficient (r): 0.95.

The four groups formed by the UPGMA method were highly influenced by fiber strength and cottonseed productivity, as they are the factors that most contributed to genetic dissimilarity (Figure 4), therefore, the greater the variation in cottonseed productivity and fiber strength, the greater the divergence between the genotypes. The most productive and strongest genotypes, BRS-Rubi (RC) and BRS-Topázio (TC) and are in the same group.
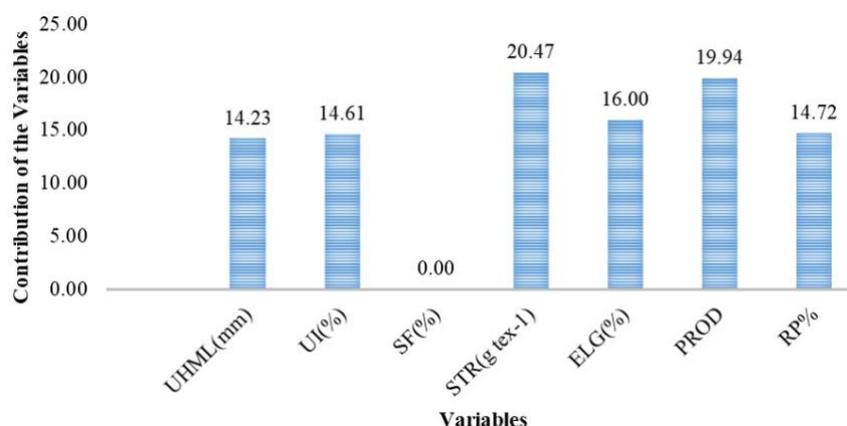
Figure 4. Relative contribution of characteristics to diversity, using the Singh method (1981), by the generalized Mahalanobis distance.

The Mahalanobis method is one of the most used in breeding to estimate dissimilarity, however for its reliability it is necessary that they have a multinormal distribution and homogeneity of the residual covariance matrix. To circumvent these limitations, computational intelligence is an alternative, as it depends only on learning and has no assumptions about the model, using a non-linear structure, such as ANN's that emulate the human brain, simulating and adjusting information by synaptic weights, similar to biological neural connections (Cruz and Nascimento 2018).

Through discriminant analysis with graphic dispersion using the ANN's, eight distinct groups were formed. The RC and TC genotypes remained isolated, as well as UFUJP-16 and UFUJP-17, similar to the dendrogram. However, the ANN's were more representative when we analyzed the other clusters as they were more coherent in relation to the means of the genotypes (Figure 5).
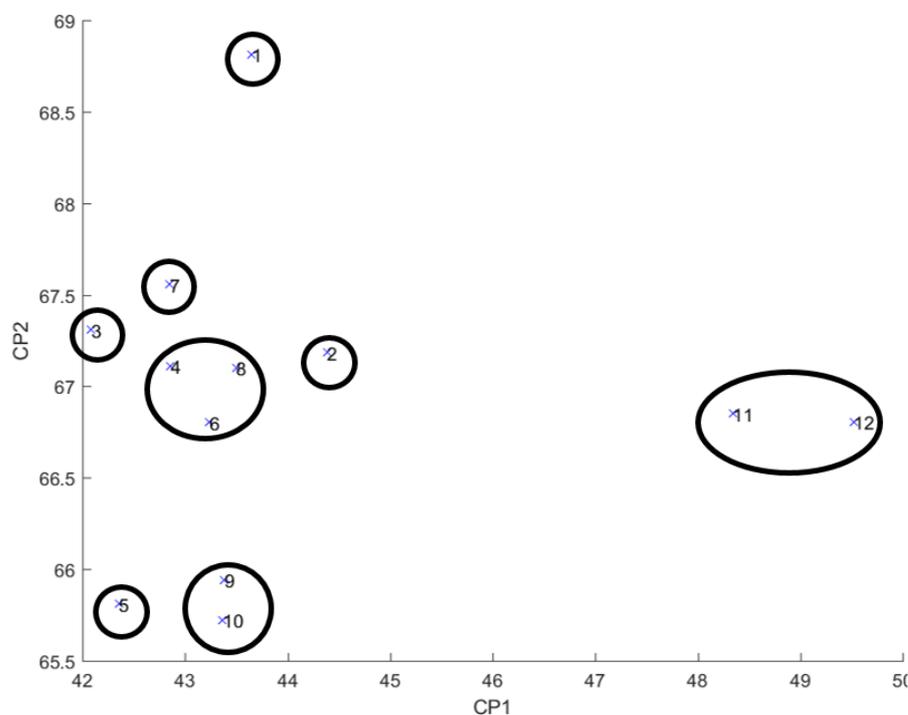
Figure 5. Graphical dispersion of scores of 12 cotton genotypes in relation to two canonical variables (CP1, CP2), based on 7 morphological characters.1 = UFUJP-01; 2 = UFUJP-02; 3 = UFUJP-05; 4 = UFUJP-08; 5 = UFUJP-09; 6 = UFUJP-10; 7 = UFUJP11; 8 = UFUJP-13; 9 = UFUJP-16; 10 = UFUJP-17; 11 = BRS RUBI; 12 = BRS TOPÁZIO.

The genotypes that showed the lowest productivity (UFUJP-01, UFUJP-05 and UFUJP-13) were allocated in different groups, which was not observed by the dendrogram (Figure 3), which suggests a different importance attributed to each method, for the characteristics. The UFUJP-05 genotype obtained the second longest elongation (8.99), UFUJP-01 one of the longest fiber lengths (24.29).

The lowest productivity was 1829.41kg ha^{-1} (UFUJP-13) (Table 1) and was grouped with UFUJP-08 and UFUJP-10. Intermediate productivity was decisive for grouping UFUJP-02 and UFUJP-09, as they obtained the fifth and sixth highest productivity, with UFUJP-02 having a high length value and a high short fiber index. The UFUJP-09 genotype has the highest IFC and the lowest fiber yield and, due to its unique characteristics, is isolated.

The greatest formation of groups in ANN's was due to the method not being affected by experimental errors, as there may be unbalanced data that do not meet the assumptions. Another relevant point is the fact that they are not based only on means and variances (Cruz and Nascimento 2018), but also increase the number of observations and decrease the apparent error rate, by quantifying the weights between neurons.

The Kohonen Self-Organizing Map (SOM) using ANN's has the ability to detect and organize the similarities of the input patterns through competitive learning, simulating the cerebral cortex with connections between the strongest neurons due to their proximity (Braga 2011; Cruz and Nascimento 2018).

Light colors show less distance between neurons which means that the characteristics have more importance, for the distinction of groups. On the other hand, dark colors represent greater distances, and therefore IFC, ALG and Pkg were the highest determining weights, respectively, in the formation of groups, corroborating the contribution of Singh (1981) only in productivity. It is possible to check the characteristics and their weights in the activation of each SOM neuron. The UHML, STR and Re% were correlated with each other, as they have the same distance pattern, represented by the same color pattern. Only the IFC and ALG characteristics did not contribute to the formation of the line 1 column 1 neuron, where RC and CT were grouped, which suggests that they were the characteristics that contributed the least to the classification (Figure 6).
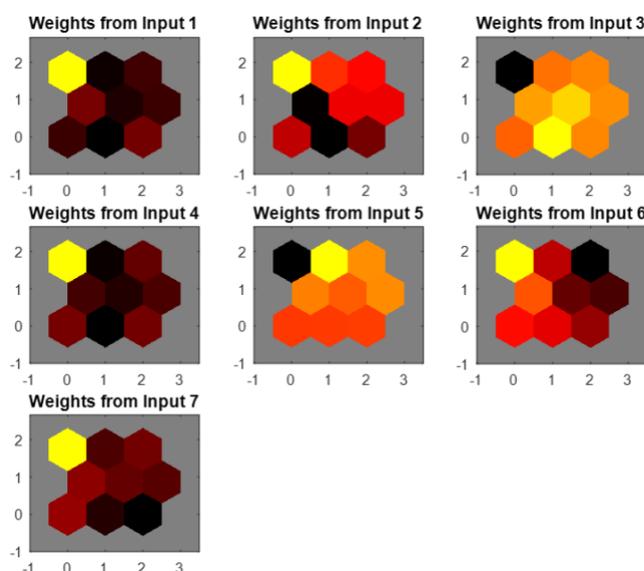
Figure 6. Influence of the weights of the input variables (mean coefficient of variation (CV) and the average sum of the seven characteristics) in the neural neurons network to define the cluster. Input 1 = UHML: fiber length; Input 2 = UI: length uniformity; Input 3 = IFC: short fiber index; Input 4 = STR: resistance; Input 5 = ALG: Elongation; Input 6 = Pkg: productivity; Input 7 = Re%: fiber yield.
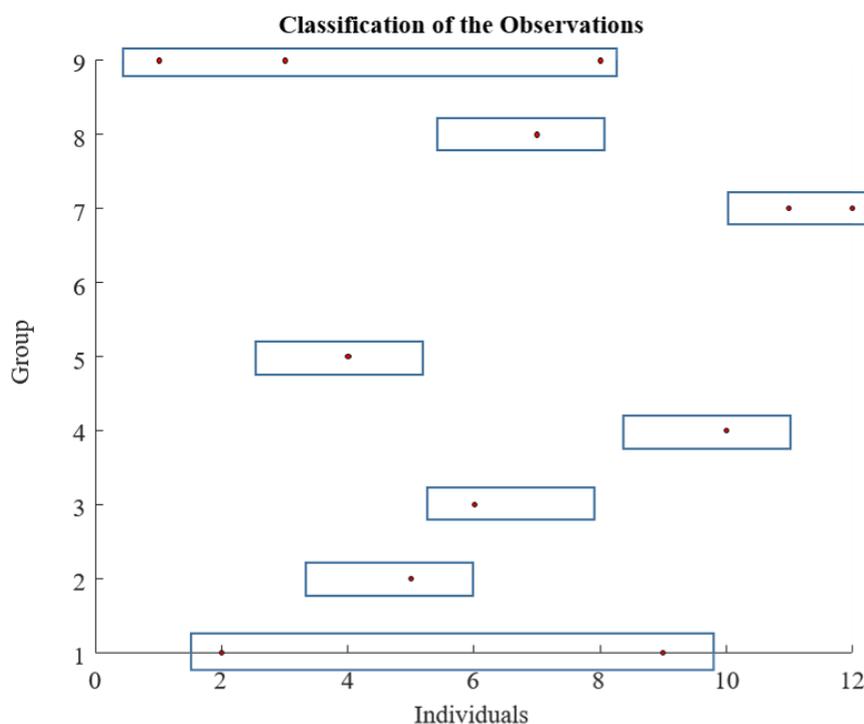

Figure 7. Kohonen's self-organizing map, with number of individuals classified into each neuron, clusters (3x3 of radius 2) using an artificial neural network.

It is possible to observe a group for the RC and TC genotypes, corroborating with the other methods (Figure 7). Distinction justified by having the highest averages for all characteristics.

There was a good representation of the SOM method. The genotypes with lower productivity were isolated (UFUJP-01, UFUJP-05 and UFUJP-13), demonstrating that this characteristic was very important in all methods to determine the classification.

The high simulation capacity of the neural networks, expand the input data by estimating new values, validating them and adjusting weights for each variable in the connections between neurons, organizing the groups by similarity through competitive learning (Cruz and Nascimento 2018) and this allows for a better distinction between genotypes (Figure 8). The neuron in row 1 columns 3 was the neuron that most grouped, these being the genotypes with the lowest productivity and low fiber quality. This grouping was distant from

the neurons that grouped the highest productivity, in the first column and with that there are less similarities between these accessions. This method distinguishes and classifies neurons (genotypes) by distance, that is, the closer the neurons, the greater the affinity between them.
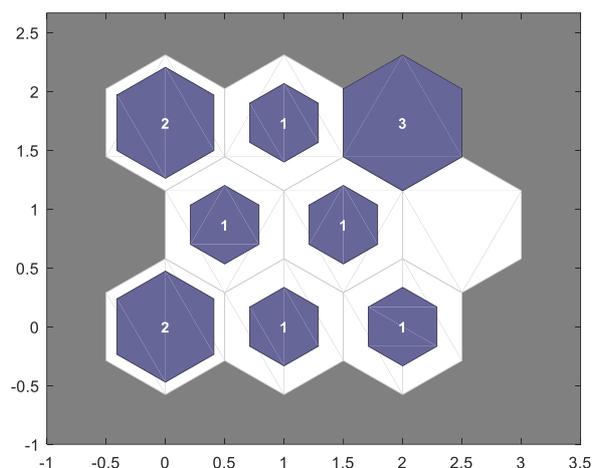
Figure 8. Topological map in the Kohonen Self-Organizing network for nine 3x3 classes of radius 2.

SOM was able to detect differences and organize similarities between accesses in a more coherent way, forming a larger number of groups, when compared to the UPGMA method and graphical dispersion, being also more accurate than the discriminant analysis, which does not corroborate Silva (2018), that when analyzing the genetic divergence between partially inbred lines of corn by multivariate methods and artificial neural networks, found greater coherence for the canonical variables.

4. Conclusions

The methods that use computational intelligence proved to be more efficient to detect similarity. Kohonen's Self-Organizing Map was the most suitable to classify and group the colored fiber cotton genotypes.

References

ASSOCIAÇÃO BRASILEIRA DOS PRODUTORES DE ALGODÃO (ABRAPA). *Algodão no mundo*. Available from: https://www.abrapa.com.br/Paginas/dados/algodao-no-mundo.aspx.

BHERING, L.L., et al. Application of neural networks to predict volume in eucalyptus. *Crop Breeding and Applied Biotechnology.* 2015, **15**, 125-131. https://doi.org/10.1590/1984-70332015v15n3a23

BORÉM, A. and FREIRE, E.C. *Algodão: do plantio a colheita*. 1th ed. Viçosa: UFV, 2014.

BRAGA, A.P., CARVALHO, A.C.L.F. and LUDEMIR, T.B. *Redes Neurais Artificias: Teoria e aplicações*. 2th ed. Rio de Janeiro: LTC, 2011.

CARDOSO, D.B.O., et al. Colored fiber cotton in the Uberlândia region using artificial neural networks for yield assessment. *Genetics and Molecular Research.* 2019, **18**(1), 13. https://doi.org/10.4238/gmr18104

CRUZ, C.D. Genes Software: extended and integrated with the R, Matlab and Selegen. *Acta Scientiarum Agronomy***.** 2016, **38**(4), 547-552. https://doi.org/10.4025/actasciagron.v38i4.32629

CRUZ, C.D., REGAZZI, A.J. and CARNEIRO, P.C.S. *Modelos biométricos aplicados ao melhoramento genético***.** 3th ed. Viçosa: UFV, 2014.

CRUZ, C. D. and NASCIMENTO, M. *Inteligência computacional aplicada ao melhoramento genético.* 1th ed. Viçosa: UFV, 2018.

DUTT, Y., et al. Breeding for high yield and quality in colored cotton. *Plant Breeding.* 2008, **123**(1), 145-151. https://doi:%2010.1046/j.1439-0523.2003.00938.x

HU, Z., ZHAO, Q. and WANG, J. The Prediction Model of Cotton Yarn Quality Based on Artificial Recurrent Neural Network. *In International Conference on Applications and Techniques in Cyber Security and Intelligence*. 2019, **1017**, 857-866. Springer, Cham. https://doi.org/10.1007/978-3-030-25128-4_105

MIRANDA, M.C.C. *Diversidade genética entre genótipos de algodoeiro visando ampliação da variabilidade.* Uberlândia: Universidade Federal de Uberlândia, 2019. Available from: https://repositorio.ufu.br/handle/123456789/24947

NARDINO, M., et al. Genetic divergence among corn (*Zea mays* L.) genotypes in distinct environments. *Revista de Ciências Agrárias*. 2017, **40**(1), 164-174. https://doi.org/10.19084/RCA16013

OLIVEIRA, A.C.L., et al. Use of mathematical modeling (artificial neural networks) in classification of banana autotetraploid (Musa acuminata colla). *Bioscience Journal*. 2013, **29**(3), 617-622, 2013.

RODRIGO, S.E., LESCANO, C.N. and RODRIGO, R.H. Application of Kohonen maps to kinetic analysis of human gait*. Revista Brasileira de Engenharia Biomédica***. 2012, **28**(3), 217-226. https://doi.org/10.4322/rbeb.2012.027

SILVA, P.C.D. *Divergência genética entre linhagens parcialmente endogâmicas de milho pipoca por métodos multivariados e redes neurais artificiais.* Uberlândia: Universidade Federal da Grande Dourados, 2018. Available from: http://repositorio.ufgd.edu.br/jspui/handle/prefix/1032