GMR
The Original by FUNPEC-RP

Support vector machines applied to the genetic classification problem of hybrid populations with high degrees of similarity

V.P. Carvalho[1], I.C. Sant'Anna[1], M. Nascimento[1], A.C.C. Nascimento[1], C.D. Cruz[2], W.A. Arbex[3], F.C. Oliveira[4], F.F. Silva[5]

[1] Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2] Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[3] Embrapa Gado e Leite, Estrada da Ribeira, km 111 – Colombo, PR – Brasil
[4] Departamento de Engenharia de Produção, Universidade Salgado de Oliveira, Juiz de Fora MG, Brasil,
[5] Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, MG, Brasil

Corresponding author: I.C. Sant'Anna
E-mail: isabelacsantanna@gmail.com

Genet. Mol. Res. 17 (4): gmr18122
Received August 22, 2018
Accepted October 26, 2018
Published October 31, 2018
DOI http://dx.doi.org/10.4238/gmr18122

ABSTRACT. Selection of appropriate genitors in breeding programs increases gains due to the variability found in the divergent groups; this allows quantification of the existing variability, saving time and resources. There are many methods for quantification and evaluation of diversity in population studies, among which we highlight methods that are based on multivariate statistical analyses, such as linear discriminant analysis (LDA) and cluster analysis. Here we propose and evaluate the use of Support Vector machine (SVM) and Artificial Neural Network (ANN) in an attempt to solve the problem of genetic classification of hybrid populations with high degrees of similarity. The results obtained, in terms of the apparent error rate (APER), were compared with those obtained using ANN analysis and LDA. In general, the lowest APER values were associated with scenarios with low degrees of genetic similarity between populations.

Specifically, the best results obtained through SVM (ranging from 14.44 to 67.41%) were observed when the exponential radial base kernel function was used. The APERs obtained by the ANN were even lower than those of the linear discriminant function.

Key words: Computational Intelligence; breeding; multivariate approach

INTRODUCTION

Genetic diversity analyses provide an opportunity for plant breeders to develop new and improved cultivars with desirable characteristics (Govindaraj et al., 2015). The selection of appropriate genitors in breeding programs has led to increased gains due to the variability found in the divergent groups. Genetic diversity studies have allowed quantification of existing variability, facilitating the management of germplasm collections, saving time and resources (Cruz et al., 2011; Sant' Anna et al., 2015). There are many methods for quantification and evaluation of diversity in population studies, among which we highlight methods that are based on multivariate statistical analyses, such as linear discriminant analysis (LDA) and cluster analysis (Berwick, 2003; Costa et al., 2006; Hamel et al., 2011).

Nogueira et al. (2008), using LDA, identified and evaluated new characteristics for production purposes and differentiation of soybean cultivars. They stated that these characteristics are useful as additional descriptors of soybean cultivars. Gonzalvez et al. (2011) successfully employed LDA to discriminate the geographical origin of rice from several PDOs (Protected Designation of Origin) recognized in Spain. Zhang et al. (2005) evaluated the potential of LDA to detect candidate markers associated with agronomic traits among 218 inbred lines of rice (*Oryza sativa*) from the United States and Asia. The results of that study lead us to suggest that LDA can be used to identify candidate markers associated with agronomic traits. However, in situations in which populations are not linearly separable, traditional methods of multivariate analysis are ineffective for classification analyses because of the difficulty of analyzing the data.

Sant'Anna et al. (2015) proposed a solution for non-linear problems; they used Artificial Neural Networks (ANNs)(Silva et al., 2010) for the genetic classification of simulated hybrid populations. They observed up to 97.5% genetic similarity, which is quite satisfactory. Another method that has been developed over the years and can be used in discrimination problems is denoted as Support Vector Machine (SVM) (Lorena et al., 2003). SVM is based on Statistical Learning Theory (Vapnik, 2013) and differs from ANNs mainly in terms of the mode of convergence; while in ANN there can be many solutions converging to local minimums, SVM converges to a single optimal solution, the global minimum (Rychestsky et al., 2001). Martins et al. (2007), in comparisons between ANNs and SVMs for the detection of leaks in oil pipelines, found that SVMs showed greater robustness and greater correct resolution of the problem.

As regards genetic improvement, SVMs have been applied in Genomic Wide Association studies (Mittag et al. 2012; Kim et al. 2013), in Genome Wide Selection (Long et al. 2011), for automating disease detection in tomato crops (Prince et al. 2015; Mokhtar et al., 2015), for predicting hybrid performance in maize crops (Lima et al., 2018), in image processing of rice (Maione and Barbosa., 2018) and for classification problems (Li and

Ogihara, 2006). In spite of the great potential for classification problem solving as an alternative to the approach based on ANNs (Sant' Anna et al., 2015). SVMs have not yet been evaluated for the purpose of solving a genetic diversity problem of hybrid populations with high degrees of genetic similarity.

In view of the above, we evaluated the use of SVMs in an attempt to solve the genetic classification problem of hybrid populations with high degrees of similarity. The results obtained, in terms of the apparent error rate (APER), were compared with those obtained using ANN analysis and Anderson's Discriminant Analysis.

Material and Methods

Simulated data

Initially, the genotypic data of nine populations in Hardy-Weinberg equilibrium (Cruz et al., 2011) with 100 plants each were simulated. Then, information on 50 codominant markers was generated and used to calculate the dissimilarity matrix by Nei's genetic distance (Nei, 1973). For each simulation, it was considered that the variance and covariance matrices would be the same for each population, since without this assumption there would be loss of linearity of the discriminant functions. Among the 10simulated populations, a pair of most divergent populations was chosen to generate hybrid F1 and two generations of hybrids (backcrossing) in relation to each parent (P_1 and P_2) (Figure 1).

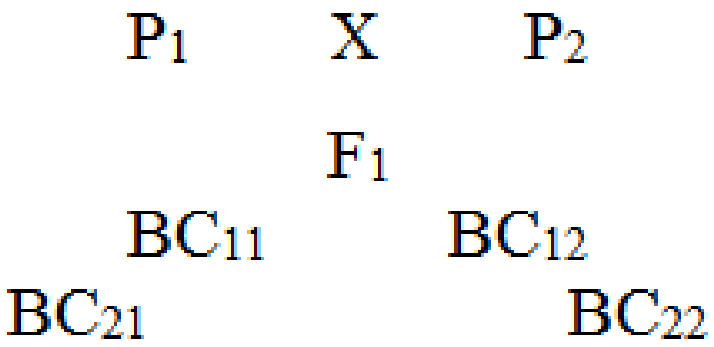

Figure 1. Structured diagram of backcrossing between P_1 and P_2 and their respective backcrosses (a) and (b).

Figure 1 shows the genotypic information for seven populations (π_j with $j = 1$ for P1; $j = 2$ for P2; $j = 3$ for F1; $j = 4,5$ for BcX; and $j = 6,7$ for BcX) that was used to simulate the phenotypic values of eight quantitative traits. Each trait was assumed to be controlled by 20 random loci, with differential additive effects determined by weights given by a binomial distribution, representing the importance of the locus in the total genotypic variability of the trait and a mean degree of dominance equal to zero. The expression $Y_{ij} = \mu + G_i + \varepsilon_{ij}$ was used, where: Y_{ij} corresponds to the phenotypic value; μ is the overall mean of the trait; G_i is the genetic effect associated with the i[th] individual from the j[th] population, given by the weighted sum of the effects of each explanatory marker of the trait; ε_{ij} is the random error, and $\varepsilon_{ij} \sim N(0, \sigma^2)$, assuming heritability (h^2) values of 55,

60, 70, ..., 90%, and numerical mean values equal to the heritability values. The genotypic, phenotypic, breeding and population simulations were performed using the simulation module of the GENES software system (Cruz, 2016). The population set was considered in three scenarios with distinct degrees of differentiation determined by the degree of similarity of the populations involved (Table 1).

Table 1: Constitution of different genetic similarity scenarios to be analyzed by the discriminant functions and Artificial Neural Networks and Support Vector Machine.

Scenarios	Populations	Similarity	Size
1	P_1, P_2, F_1	50%	300
2	P_1, P_2, F_1, Bc_{11}, Bc_{12}	75%	500
3	P_1, P_2, F_1, Bc_{11}, Bc_{21}, Bc_{12}, Bc_{22}	87.5%	700

Discrimination Methods

Discriminant Analysis

For the estimation of the discriminant functions used to classify the groups of individuals, π_g is considered as the populations to be compared, given μ_g and Σ_g – the mean vector and the covariance matrix of these populations, respectively –, where g={1, 2,..., 7} and *n* varies from 3 to 5 and to 7, given that the scenario consists of populations of three to seven groups. The simulation considered populations of the same variance, so $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma_p$, where Σ_p is the common covariance matrix given by equation

$$\Sigma_p = \frac{\sum_{i=1}^{g}(n_i-1)\Sigma_i}{\sum_{i=1}^{g}(n_i-1)} \quad \text{Equation (1)}$$

and the discriminant function given by the following expression:

$$D_i(X) = \ln(p_i) + \mu_i^T \Sigma^{-1} X - \frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i \quad \text{Equation (2)}$$

for $i = 1, \ldots, n$ [9] in which p_g is the a priori probability that group g will belong to population π_g. The new individual is classified as belonging to the group with the highest classification score; an individual X will be classified in group π_g if: $D_g(X_k) = \max(D_1(X), \ldots, D_n(X))$.

Artificial Neural Network (ANN)

Artificial neural networks are biologically inspired computer programs designed to simulate the way in which the human brain processes information. The ANN is formed by a combination of several artificial neurons, able to simulate the behavior and functions of a biological neuron. The artificial neuron receives one or more inputs and sums them to produce an output; therefore, each neuron represents an output (Braga et al., 2011). As regards biological systems, these connections represent the contacts of the dendrites with other neurons, thus forming synapses (communication between two cells). Such connections make the output signal of a neuron an input signal of another. Here, we used the networks known as multilayer perceptron, also known as feed forward networks (Silva et al., 2010). In such networks, the information flow propagates forward, layer by layer, from the input

layer to the output layer, with one or more layers of neurons between the input and output layers. Figure 2 shows how neurons are distributed in a multilayer perceptron (Sant'Anna et al., 2015).

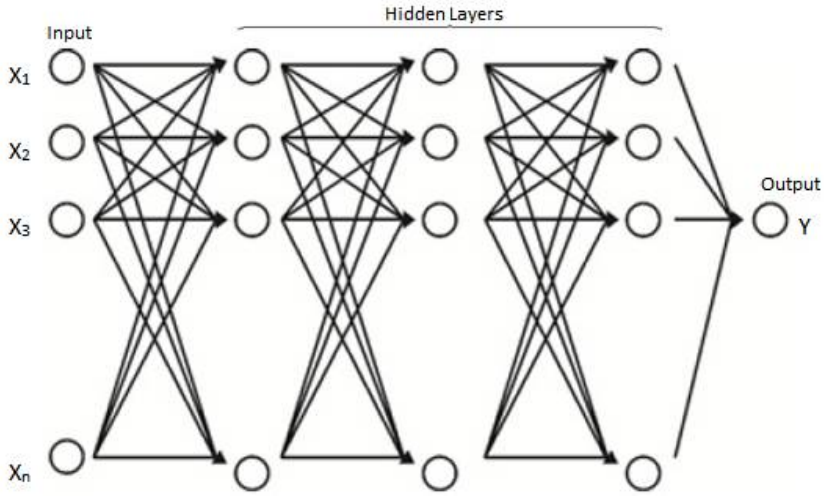

Figure 2: Representation of a neural network model Multilayer Perceptron used for classification. Source: Adapted from Sant'Anna et al. (2015).

The *backpropagation* algorithm Silva et al. (2010) was used in the training process. The neural network architecture (Figure 2) consisted of three hidden layers, tansig or logsig activation functions, number of neurons varying from 6 to 15 neurons in the first, 10 to 40 in the second, and 10 to 40 in the third layer, as suggested by Sant'anna et al (2015). The minimum number of iterations (or epochs) was 1500 and all combinations of neuron numbers and activation functions in the hidden layers were checked; 70% of the data was used as a training set and the test set was composed of the remaining 30%.

Support Vector Machine

The proposed method known as Support Vector Machine (SVM) (James et al., 2013) is based on statistical learning theory, which aims to establish mathematical conditions that allow choosing a classifier with good performance from the set of data available for training and testing (James et al., 2013). The objective is to provide a classifier that shows good performance for the samples that were not observed during the training. This is done using the method of structural risk minimization, which depends on a term called Vapnik-Chervonenkis dimension (Vapnik, 2013) that measures the intrinsic complexity of a class of functions. The main idea of SVM is to create a separating hyperplane as a decision surface so that the separation between its positive and negative examples is maximal (James et al., 2013; Haykin, 2009; Campbel, 2000).

In this way, a mathematical model is defined, as follows:

$$w \cdot x + b = 0, \qquad \text{Equation (3)}$$

where w is the adjustable weight vector and b is a bias. Thus, the space of data X is divided into two regions: $w \cdot x_i + b > 0$ and $w \cdot x_i + b < 0$ such that $g(x) = sgn(f(x)) = sgn(w \cdot x + b)$ so that the classification will be +1 if $f(x) > 0$ and -1 if $f(x) < 0$, and the samples that serve as the basis for modeling the separation margin are called support vectors (Campbel, 2000). Accordingly, it is by means of support vectors that one can tell if a pattern belongs or not to a certain class, depending on the value obtained in $f(x)$, as can be seen in Figure 3.

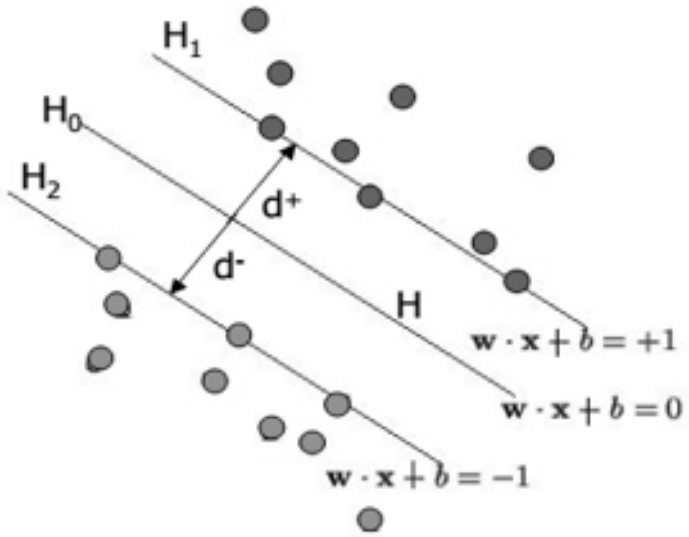

Figure 3- Linearly separable data set with the separating hyperplane and the separation margin.

In Figure 3, d_+ represents the shortest distance from the nearest positive point and d_- is the shortest distance from the nearest negative point (support vectors). However, the data in this paper do not present a linear structure, so a separating hyperplane can be defined as follows:

$$f(x) = w \cdot \varphi(x) + b, \qquad \text{Equation (4)}$$

where $\varphi(x)$ is known as a kernel. This function aims to map the training data set into higher dimensionality spaces, which leads to a linear classification problem. Another way to deal with the nonlinearity of the data is to implement a smoothing constant (C), which determines the rigidity of the separation margin. One must be very careful with its use, because for a given optimal value (C), depending on how far this value is from the observed data, generalization can be compromised and inform that the data samples are equivalent to the support vector points (Matlab, 2011). We used the polynomial kernel function (PolyKernel), the radial basis function (Normalized PolyKernel) and the exponential radial basis function (RBFPolyKernel). The functions and the search spaces of the parameters analyzed are displayed in Table 2. The search space of the smoothing constant (C) was $1 \leq C \leq 100$, with variation of one in each loop for all functions used with

the same criteria. Finally, in order to separate the population groups, as more than three groups were to be classified, the one-against-all space decomposition approach was used in the set of variables.

Table 2. Choice of parameters for different kernels.

Kernel	Function	Rangeof values	Variation of values
Gaussian (RBF)	$exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$	$0.001 \leq \frac{1}{2\sigma^2} \leq 0.5$	0.0001
Polynomial	$\gamma(x_i \cdot x_j + \alpha)^d$	$1 \leq d \leq 10$ $0.001 \leq \gamma \leq 0.2$ $-1 \leq \alpha \leq 1$	1 0.001 0.1
Sigmoidal	$tanh(\beta_0(x_i \cdot x_j) + \beta_1)$	$0.001 \leq \beta_0 \leq 0.5$ $0.001 \leq \beta_1 \leq 0.5$	0.001 0.001

As in the ANN analyses, the data were separated in a way that one part was for training and the other for validation.

Comparison between methods

The APER was given by the ratio between the number of erroneous classifications and the total number of classifications (Cruz et al., 2014), according to: $AER(\%) = \frac{1}{N}\sum_{j=1}^{k} m_j$, where m_j is the number of observations of population π_j, which were, by means of discriminant functions, classified in another population $\pi_{j'}$, where $j' = j$ and $j = 1, 2, ...,$ 7 populations; considering: $N = \sum_{j=1}^{100} n_j$, where n_j is the number of observations related to population π_j.

Computational aspects

Simulation processes were run using the simulation module of the GENES software system Cruz, 2016). ANN was implemented for the integration between Genes and MATLAB (Cruz, 2016; Matlab, 2011). Discriminant analysis was implemented using ksvm functions, and SVM predictions were performed using the package kernlab (Karatzoglou, 2018) from software R (R Core Team, 2018).

RESULTS

The APER values, for the results obtained by LDA and SVM, ranged from 14.44 (~ 101 individuals) to 67.14% (~ 470 individuals). In general, the lowest APER values were associated with scenarios with low degrees of similarity between populations (Table 3). For instance, APER was equal to 18.89 and 14.44%, considering the results obtained by LDA and SVM in Scenario 1 (P1, P2 and F1). Specifically, the best results obtained through SVM (those ranging from 14.44 to 67.41%) were observed when the exponential radial base kernel function was used (Table 3). The APERs obtained by the ANN were even lower than those of the discriminant function.

Table 3. Apparent Error Rate (APER) in percent estimated by Anderson's Discriminant Analysis, Support Vector Machine and Artificial Neural Networks applied in all three scenarios distinct degrees of similarity of the populations involved where 1=(P_1, P_2, F_1), 2=(P_1, P_2, F_1, Bc_{11}, Bc_{12}) and 3=(P_1, P_2, F_1, Bc_{11}, Bc_{12}, Bc_{21}, Bc_{22}).

Scenarios	LDA	SMV			ANN
		RBF	Polynomial	Sigmoidal	MLP
1	18.89	22.22	17.78	14.44	77.14
2	52.00	54.67	51.33	51.33	73.57
3	67.14	61.61	61.61	63.03	71.42

The parameters that showed the best results in Scenario 1 were, respectively, 11, 0.102 and 0.028 for the smoothing constant (C) and parameters β_0 and β_1 of the hyperbolic tangent function. The smoothing constant was small (1 or 3) for the other scenarios, with parameters β_0 and β_1 showing divergent results between scenarios.

Table 4. Support Vector Machine in all three scenarios of genetics similarity (50, 75 and 87.5%) for different values of the smoothing constant (C) and optimal parameters.

Scenarios	RBF		Polynomial				Sigmoidal		
	C	$1/2\sigma^2$	C	d	γ	α	C	β_0	β_1
1	1	0.002	1	5	0.040	-0.4	11	0.102	0.028
2	1	0.387	3	3	0.032	-0.9	1	0.050	0.258
3	11	0.088	3	6	0.029	1	1	0.183	0.400

Different from the results obtained through DA and SVM, the APER values considering ANN were zero in all scenarios (Table 3). The number of neurons used to solve the problem ranged from 6 to 40 (Table 5). In general, the scenario with greatest similarity (87.5%) among the populations demanded a greater number of neurons. Specifically, 15, 30 and 40 neurons were required respectively in layers 1, 2 and 3 for the solution of the problem, as for Scenario 3 (Table 5). For that scenario, the Log-sigmoid function (logsig) and the hyperbolic tangent function (tansig) were used to obtain the solution to the discrimination problem.

Table 5. Neural network topology. Numbers of neurons per layer (L1, L2, L3) and activation functions (F1, F2, F3) for each layer in the three scenarios of genetic similarity (50, 75and 87.5%).

Scenarios	Number of Neurons			Activation Function		
	L1	L2	L3	F1	F2	F3
1	6	10	15	tansig	tansig	logsig
2	6	30	30	tansig	tansig	tansig
3	15	30	40	logsig	tansig	logsig

DISCUSSION

We propose the use of the SVM method as a computational tool to obtain a solution to the genetic classification problem of backcross populations with high degrees of similarity. Comparisons between the SVM, LDA and the ANN methods were made under different simulated scenarios. The scenarios consisted of three groups of populations with different degrees of similarity (L1=50.00; L2=75.00; L3=87.50). Predictive performance was measured with the apparent error rate (APER).

The use of a Support Vector Machine was efficient in obtaining a solution to the genetic classification problem of hybrid populations with high degrees of similarity, since the APER values were similar or lower than those obtained by applying LDA. LDA is

suitable for situations in which populations are linearly separable (Li et al. 2006). Li et al. (2006) investigated the use of LDA and SVM for multi-class classification. Their experiments showed that the precision of the LDA approach is comparable to that of other approaches, such as SVM. Moreover, the results found using LDA were expected, since the genetic structure imposed in the simulation process showed similarity between the populations, ranging between 50 (L_1) and 87.5% (L_3). In the study carried out by Sant'Anna et al. (2015), also using LDA to classify hybrid populations, the APER values ranged from 22.67 to 80.01%, being similar to those we obtained here. Nevertheless, LDA is still very useful for breeding programs in simpler scenarios and has been successfully used in some crops with the objective of generating discriminant functions such as in bean genotypes (Li et al., 2006), for differentiating soybean cultivars (Nogueira et al., 2008) and for distinguishing rice cultivars (Maione and Barbosa, 2018).

The ANN analyses gave unsatisfactory results in terms of a solution to the classification problem in populations with high degrees of similarity. The APER values were higher than 70%, being different from those obtained in Sant'Anna's study (Sant' Anna et al., 2015), where ANN and LDA were compared in a classification problem of backcross populations. According to Braga et al. (2011), ANNs are superior to the conventional methods most commonly used for population discrimination. One complexity related to the use of ANNs is the definition of the number of hidden layers, neurons and activation functions used in the ANN architecture. In our study, this problem was minimized as these parameters were chosen considering the values used by Sant'Anna et al. (2015), who used three hidden layers with 15, 40 and 40 neurons in the first, second and third hidden layers, respectively. This fact emphasizes the importance of studies involving computational intelligence techniques. However, even using similar parameters, the neural network used here was not the same as that in the above-mentioned study; the weights used in the two networks were different as well as the populations used in their training. This suggests that it is necessary to test more algorithms in different population groups and use different forms of validation so that the reliability of the results can progressively improve.

Although SVMs showed a learning capacity, such as the perceptron (Lorena and Carvalho 2003; Martins, 2007), we obtained different results. Possibly, owing to the high complexity of the problem under study, it will be necessary to make the parameter space even more flexible. To this end, computational techniques are required to optimize the search in the parameter space. Among these techniques, those that reduce dimensionality can be considered, in addition to other types of kernels.

ACKNOWLEDGMENTS

We acknowledge the Foundation for Support of the Federal University of Viçosa (FUNARBE), the Foundation for Research Support of Minas Gerais State (Fapemig - PPM-00518-15), the Brazilian Federal Agency for Support and Evaluation of Graduate Education (Capes), and the National Council for Scientific and Technological Development (CNPq) for financial support.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Berwick R (2003). *An Idiot's Guide to Support Vector Machines*. Advanced Computer Vision. University of Central Florida, Orlando.

Braga AD, Carvalho APLF, Ludermir TB (2011). *Redes neurais artificiais: teoria e aplicações.* 2ª ed., LTC, Rio de Janeiro.

Campbell C (2000). An introduction to kernel methods. In Howlett, RJ and Jain, L C, editors, *Radial Basis Function Networks: Design and Applications.* SpringerVerlag, Berlim.

Costa MD, Pereira WE, Bruno RDL, Freire EC, et al.,.(2006). Divergência genética entre acessos e cultivares de mamoneira por meio de estatística multivariada. *Pesq. agropec. bras.*41:1617-1622.

Cruz CD, Ferreira FM, Pessoni, LA. (2011). *Biometria aplicada ao estudo da diversidade genética.*Suprema. Visconde do Rio Branco.

Cruz CD (2016). Genes Software – extended and integrated with the R, Matlab and Selegen. *Acta Sci.-Agron.*38:547-552.

Gonzalvez A, Armenta S, Guardia M DE LA (2011). Geographical traceability of Arròs de Valencia rice grain based on mineral element composition. *Food Chem*. 126:1254–1260.

Govindaraj, Mahalingam, M. Vetriventhan, M. Srinivasan. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: an overview of its analytical perspectives.*Genet. res. int.* Available at [https://www.hindawi.com/journals/gri/2015/431487/]

Hamel, L H. (2011). Knowledge discovery with support vector machines. *John Wiley & Sons*. New Jersey.

James G, Witten D, Hastie T,Tibshirani R. (2013). *An introduction to statistical learning*. Springer, New York.

Karatzoglou A, Smola A, Hornik K, Karatzoglou MA. (2018) Package 'kernlab'.

Kim J, Sohn I, Kim DDH, Jung SH. (2013). SNP selection in genome-wide association studies via penalized support vector machine with MAX test. *Computational and mathematical methods in medicine*. Article ID 340678, 8 pages. Avaialable:[https://www.hindawi.com/journals/cmmm/2013/340678/]

Li T, Zhu S, Ogihara M. (2006). Using discriminant analysis for multi-class classification: an experimental investigation. *Knowl. Inf. Syst., 10*:453-472.

Lima, FDA., Willmitzer L, Nikoloski Z(2018). Classification-driven framework to predict maize hybrid field performance from metabolic profiles of young parental roots. *PloS one, 13*:e0196038.

Long, N, Gianola D, Rosa GJ, Weigel KA. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet., 123*:1065.

Lorena AC, Carvalho AC. (2003). *Introdução as máquinas de vetores suporte*. Relatório Técnico do Instituto de Ciências Matemáticas e de Computaçao. USP, São Carlos.

Maione C, Barbosa RM, (2018). Recent applications of multivariate data analysis methods in the authentication of rice and the most analyzed parameters: A review. *Critical reviews in food science and nutrition*, pp.1-12.

Martins RS, Duarte VJ L, Maitelli, AL, Salazar AOet al.,. (2007). Sistemas de Detecção de Vazamentos em Dutos Usando Redes Neurais e Máquinas de Vetor de Suporte. Anais do VIII Congresso Brasileiro de Redes Neurais, pp. 1-6, Florianópolis.

MATLAB version .7.12.0.635 (2011). Natick, Massachusetts: The MathWorks Inc., 2010.

Mittag F, Büchel F, Saad M, Jahn A,et al. (2012). Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. *Hum. mutat*, 33:1708-1718.

Mokhtar U, Ali Ma, Hassanien, AE, Hefny H, (2015). Identifying two of tomatoes leaf viruses using support vector machine. In *Information Systems Design and Intelligent Applications* (771-782). Springer, New Delhi.

Nei, M. Genetic distance between populations. (1973) *Am. Naturalist*, 106: 283-292.

Nogueira APO, Sediyama T, Cruz CD, Reis MS, et al., (2008). Novas características para diferenciação de cultivares de soja pela análise discriminante. *Cienc. Rural*, 38:1-7.

Prince G, Clarkson JP, Rajpoot NM. (2015). Automatic detection of diseased tomato plants using thermal and stereo visible light images. *PloS one*, 10:e0123262.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rychetsky, M. (2001*). Algorithms and Architectures for Machine Learning Based on Regularized Neural Networks and Support Vector Approaches*. Berichte Aus Der Informatik. *Shaker Verlag GmbH*. Germany.

Sant'anna IC, Tomaz RS, Silva GN, Nascimento M, et al.,. (2015). Superiority of artificial neural networks for a genetic classification procedure. *Genet Mol Res*, 14: 9898-9906.

Silva ID, Spatti DH, Flauzino, RA. (2010). *Redes neurais artificiais para engenharia e ciências aplicadas*. Artiliber,São Paulo.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science and Business Media, New York.

Zhang N, Xu Y, Akash M, McCouch S, et al., (2005). Identification of candidate markers associated with agronomic traits in rice using discriminant analysis. *Theor. Appl. Genet.* 110:721-729.