**Pisum sativum** classification based on a methodological approach for pattern recognition using discriminant analysis and neural networks

Anna Pérez-Méndez*, Ronald Maldonado-Rodríguez**, Elizabeth Torres-Rivas*,  
Francklin Rivas-Echeverría^  
Universidad de Los Andes, Mérida, Venezuela 5101  
*Escuela de Estadística, ^Laboratorio de Sistemas Inteligentes  
**Bioenergetic Laboratory. University of Geneva, Switzerland

**Abstract:** In this work a statistical analysis-based methodological approach for a pattern recognition system using discriminant analysis and neural networks is used for the classification of **Pisum sativum** (peas) according to the drought resistance. The statistical techniques used in the exploratory analysis are a fundamental tool in the creation of variables sets and observations for the model adjustment in the neural models and in the discriminant models.

**Key Words:** Classification, Discriminant analysis, Artificial Neural Networks, **Pisum sativum**, Statistical Analysis.

1 Introduction

One of the main areas in which it has been widely used the patterns recognition and classification, is in plants taxonomy [8]. The researchers in these areas are not only interested in selecting and classifying a certain specimen, but also they are interested in methods that can help the plants state of health recognition. One of the used methods is the JIP TEST [8], which is a tool for the plants vitality supervision.

In this work the application of the methodological approach presented in [13] is used for the development of a pattern recognition system for the classification of **Pisum sativum** (peas) according to the drought resistance. The investigation was based on the work given by Ronald Maldonado-Rodriguez in the Bioenergetic Laboratory of the University of Geneva, Switzerland [8]. In that work, it was desired to classify eight varieties of peas (**Pisum sativum**) according to the drought resistance (hydric stress) within three classes: high resistance, intermediate resistance and low resistance. In order to make the classification, two techniques for the pattern recognition are used: discriminant analysis and artificial neural networks.

The present work is organized as follows: in section 2 fundamental aspects of the methodology for patterns recognition systems using discriminant analysis and neural networks is presented and in section 3 the application of the methodology is used for the **Pisum sativum** classification problem according to the drought resistance using discriminant analysis and artificial neural networks. Section 4 presents the corresponding conclusions and recommendations.

2 Methodology for Pattern Recognition and Classification Using Discriminant Analysis and Artificial Neural Networks

This methodology presented in [13] is divided in stages and phases, which are summarized next:

Stage 1. Analysis and Description of the Problem: It is analyzed the nature and characteristics of the problem, including the study of the information sources, data infrastructure, etc.

Stage 2. Feasibility analysis for classification using Discriminant Analysis and Neural Networks: It is studied the feasibility for solving the classification problem considering the exposition made in stage 1.

Stage 3. Analysis of the Variables that take part in the Process: All the variables that take part in the process and which affect direct or indirectly on the classification variable are statistically studied in detail. This stage has three important phases: **Matrix of Data Description, Software Requirements and Exploratory Data Analysis**.

Stage 4. Input Data Requirements: The selection of the measurements and variables that will be used for constructing the discriminant and neural models is made; also, it will be made the processing and depuration of such measurements by means of statistical techniques in order to solve the problems
detected in the previous stage. This stage involves two phases: Data processing and Adjustment Set Selection.

Stage 5. Discriminant Analysis: It is made the adjustment and evaluation of discriminant models. This stage contemplates two important phases: Discriminant Models Construction and Discriminant Models Evaluation.

Stage 6. Neural Models Construction: It is given the neural networks training, whose inputs were selected in stage 4, and correspond to the training set. This stage has two important phases: Neural Networks Training and Neural networks model or models Evaluation.

Stage 7. Final Results and Conclusions: In this stage it is compared the obtained classification and the error of the best selected discriminant and neural models in stages 5 and 6 respectively, with the purpose of choosing the model that better represents the phenomenon in study.

Stage 8. Maintenance and Update of the Selected Model for the System classification: This stage must last during the system’s life, incorporating the appropriate knowledge and/or resources according to the technological requirements for its use.

3 Pisum sativum classification according to the drought resistance using discriminant analysis and artificial neural networks

3.1 Stage 1. Analysis and Description of the Problem

Plants during their growing are exposed to constantly changing conditions in the environment that surround them. When those conditions are not favorable, plant responds in negative form to these changes; this is, environmental alterations affect optimal operation and the general development of the plant. This type of alterations that negatively affect the plants physiological processes, causes that the plants present a biological stress condition [8, 17, 18].

The stress of plants takes place when they are in adverse situations, which can be originated by alive agents (biotic stress) and non alive (abiotic stress).

The answers of plants to stress, can be divided in four stages:

1. Reaction or Alarm Phase: where the affected function varies much of the normal one.
2. Resistance Phase: the plant is adapted to the stressing factor, being able to return to its normal state.
3. Exhaustion Phase: if the stress factor stays for a long time or increases in intensity, the physiological operation of the plant can vary much, and even produce the death of the vegetable.
4. Regeneration Phase: in this phase, the plant can return to its normal physiological condition, as long as the damage has not generated too many consequences.

Two important types of stress in the plants exist: hydric stress and saline stress.

hydric stress: The most important resource in plants development, is without a doubt the water. Hydric stress can be consider as a general pathological state of plant, in which the photosynthesis process gets lower and can lack the carbonic anhydrid (CO2); there exist chlorophyll loss; the enzymes become denaturalized, and the protein synthesis stops; the plant dehydrates what makes stop the growth, etc.

saline stress: The mechanisms by which the plants tolerate the salinity are complex. They involve molecular syntheses, enzyme induction and membrane transport. Saline stress takes place by the toxicity of the ground, that is, an excess of salts in the environment can induce a physiological stress in some plants. Some of the ions that more problems induce are: molecular chlorine (CL), ionic sodium (Na+), nitrate ion (NO3), sulphate ion (SO4) and ammonium ion (NH4+) [10, 18].

Actually, there exists an increasing necessity of fast methods that can help the recognition of the plants health state and that also can be useful in plants selection and classification tasks. One of the methods used for this mission is the JIP TEST [8], which, is a tool for the supervision of the plants vitality. This test is based on the fluorescence rays and multiphase chlorophyll analysis. In recently developed works in the area, it has been demonstrated that the fluorescence signals can be incorporated as a bar code of the physiological characteristics of the plants and can be used for taxonomic intentions. Evidence of it, is the experimental work of Dr Ronald Maldonado Rodriguez, in the Bioenergetic Laboratory - University of Geneva, Switzerland whose primary target is “proposing a fast method for developing
strategies that help to the diminution of the impact of environmental stress in agriculture” [8].

In the work developed by Dr Maldonado Rodriguez, it was desired to classify eight varieties of peas (*Pisum sativum*) according to the drought resistance (hydric stress) within three classes: high resistance, intermediate resistance and low or sensible resistance. In order to make the classification, fluorescence induction curves are used, which are obtained *in-vivo* and *in-situ* by means of portable fluorimeter. The experiments for making these measurements took several years. The eight varieties classification according to the drought resistance, was based on measurements of osmotic pressure and water content of the varieties used in the experiment. Of the eight varieties, three correspond to the low resistance class, two of remaining species to the intermediate resistance class and the last three species to the high resistance class.

The conditions used for the experiment are the following ones:
- 8 plants of each variety were cultivated in 15 centimeters of diameter pots in the laboratory.
- Greenhouse conditions (23/16 °C, day/night, in 14/10 hours photoperiod, HFI Hg 400W lamps).
- the plants were watered every two days with tap water.
- the measured signals of fluorescence (in miliwats/miliseconds) are obtained from the superior and inferior leaves of the plants.
- a curve for each leaf is obtained.
- There are extracted 20 values from each curve, which are used for defining the classification models input vectors.
- the 20 vector components, that defines each measured curve, are enough representative for characterizing the measured signals.

3.2 Stage 2. Feasibility analysis for classification using Discriminant Analysis and Neural Networks.

Due to the nature of *Pisum sativum* classification problem, the output variable represents three drought resistance classes that are mutually excluding and exhaustive. The independent variables are quantitative; so they present functional characteristics that allow classification rules derived from discriminant analysis and artificial neural networks. There is reliable information available, then, it is possible to determinate that it is feasible to develop a classification procedure using discriminant analysis and neural networks.

3.3 Stage 3. Analysis of the Variables that take part in the Process.

3.3.1 Phase 3.1 Matrix of Data Description

It is received a matrix of data with the following characteristics: 144 rows that correspond to the leaves observations. 21 columns, 20 correspond to the fluorescence curve values and the last column indicates the resistance class (dependents and independent variables respectively).

These data correspond only to low leaves of the plants, since the measurements tend to be more homogenous than in the high leaves, this fact is justified in order to control a variation source attributed to the foliar development [8].

The number of leaves and the corresponding percentage for each type of resistance are the following:

<table>
<thead>
<tr>
<th>Resistance</th>
<th>Number of leaves</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>57</td>
<td>39.55%</td>
</tr>
<tr>
<td>Intermediate</td>
<td>35</td>
<td>24.63%</td>
</tr>
<tr>
<td>High</td>
<td>52</td>
<td>35.82%</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 1. Available observations Number of Data corresponding to each class of drought resistance

3.3.2 Phase 3.2 Software Requirements

It must be specified the software requirements for the data statistical manipulation and for the discriminant and neural models adjustment. The software specifications are the following ones:
- S-Plus version 6.0 [19] is selected as statistical software (exploratory analysis and discriminant models)
- Statistica Neural Networks [15] is selected for training the feedforward neural networks using backpropagation algorithm.

3.3.3 Phase 3.3 Exploratory Data Analysis

The 20 values of the fluorescence induction curve have the same measurement unit: miliwats/miliseconds (mW/ms). Descriptive statistics by drought resistance group was calculated and did not appear missing values.

A important issue found in all three resistance groups, was the observation of identical information repeated by diverse values in the fluorescence induction curve; presenting the same central tendency indices, forms and dispersion, which is
considered a serious multicolinearity problem. Six (6) of the curve values are eliminated, and the matrix of data presents new dimensions.

When examining the correlation matrix, two great groups of variables, in which direct or positive linear correlation between the variables that compose them, were detected; with the particularity that both groups are complementary. Very high coefficients within each group, but low correlation coefficients between groups exist.

With all the arguments exposed until now, it is important to make a principal components analysis, with the purpose of suggesting the independent variables that will be used for constructing the discriminant and neuronal models.

In the principal components analysis, it was obtained that the first component explains 66.05% of the observations total variation, and with two first components 93.28% of the total data variation is explained, which represents a very good data explanation. The first component coefficients show a weight of all the variables, adding that all the coefficients are positive. The second component opposes the variables of group 1 with the variables of group 2.

Once reviewed the descriptive statistics, the correlation matrix and principal components, 4 independent variables are suggested, that correspond to 4 values of the fluorescence induction curve, which also represent the minimum, maximum and intermediate values of the curve (units, tens, hundreds and thousands mW/ms).

### 3.4 Stage 4. Input Data Requirements

#### 3.4.1 Phase 4.1 Data processing

In this phase it is made the processing of the four values in the fluorescence induction curve selected as independent or explanatory variables for constructing the discriminant and neural models. The processing of these values, consists of verifying, by means of graphical techniques and the treatment of these, if atypical observations exist, to verify if it is necessary to apply simple mathematical transformations that promote normality and symmetry, tests of normality, etc. [2, 4, 6, 7].

In any of the 4 values, atypical observations were detected and according to the normality Kolgomorov–Smirnov test, the distribution of these four values is normal. Mathematical transformations were not required.

### 3.4.2 Phase 4.2 Adjustment Set Selection

For constructing the discriminant and neural models, it is necessary to make a partition of the original data set of in two sets: a set for adjusting the models and the other for testing these models. In this particular application, it is necessary to make use of the stratification concept, in which the stratus are represented by the three classes of drought resistance. In order to select the sample, the proportional fixing criterion is used, in which, in the random sample, the class of resistance size is proportional to the size of each one of these groups.

In this application, 85% of the data will be used (approximately 122 observations) for constructing discriminant and neuronal models and 15% of the data (22 observations) for testing these models.

Two random stratified samples of 122 observations are selected, each one with the 4 independent selected values (selected in stage 3) of the fluorescence curve, and another stratified random sample is selected that contains 122 observations and the 14 values of the curve.

<table>
<thead>
<tr>
<th>Resistance</th>
<th>Adjustment</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>44</td>
<td>8</td>
<td>52</td>
</tr>
<tr>
<td>Intermediate</td>
<td>30</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>Low</td>
<td>48</td>
<td>9</td>
<td>57</td>
</tr>
<tr>
<td>Total</td>
<td>122</td>
<td>22</td>
<td>144</td>
</tr>
</tbody>
</table>

Table 2. Number of observations by Resistance Class of in the Adjustment Set and the Testing Set

### 3.5 Stage 5. Discriminate Analysis

#### 3.5.1 Phase 5.1 Discriminate Models Construction

It is considered, in first place, to fit linear discriminant models, that are the simplest. In this type of analysis the population variance and covariance matrices are considered equal for the three resistance groups, for checking the fulfillment of the homocedasticity assumption, the Box M Test is made, which reveals that the assumption of homocedasticity between the three classes of drought resistance is not fulfilled, so it is not appropriate to fit linear discriminant models, but it can be done fitting quadratic discriminant models, in which it is allowed that the population covariance matrices are unequal.

<table>
<thead>
<tr>
<th>Statistical</th>
<th>DF</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Box M Test</td>
<td>97.58551</td>
<td>20</td>
</tr>
<tr>
<td>adj M Test</td>
<td>92.64719</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 3. Covariance Matrices Homogeneity Test
It is adjusted quadratic discriminant models, in which it is considered that the structure of the covariance matrix is heterocedastic, and a priori probabilities are considered proportional to the size of each class of resistance. The T2* Hotelling test was made according to the approach propose by Yao [20, 21] for detecting differences between pairs of averages vectors, and in the three samples it concludes that there exist differences between the three pairs of averages vectors in at least one of their components.

With the quadratic discriminant functions, the high and intermediate resistance classes were separated good enough, but the observations that correspond to the low resistance are mixed with the high class resistance and the intermediate resistance observations, presenting greater confusion proportion with the last ones. This issue can be seen in figure 1.

![Quadratic Discriminant Function Biplot](image)

**Figure 1. Quadratic discriminant function Biplot**

### 3.5.2 Phase 5.2 Discriminate Models Evaluation

For each of the adjustment sets used for the construction of the discriminant models, it has its corresponding testing set, that consists, in every case, of a 22 observations set distributed proportionally to the size of each resistance class. The purpose of the test set, is to evaluate the model assignment to new observations, and based on the correctly classified cases proportion then decide which model is the best one, and suggest it for the new observations classification.

Next, the assignment results for new observations are presented using the discriminant quadratic functions in the three testing sets.

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Interim.</th>
<th>High</th>
<th>Total</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>9</td>
<td>1.000000</td>
</tr>
<tr>
<td>Interim.</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0.000000</td>
</tr>
<tr>
<td>High</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>1.000000</td>
</tr>
</tbody>
</table>

Total Error = 0.772727

Table 4. Testing set 1 with 4 input variables

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Interim.</th>
<th>High</th>
<th>Total</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>9</td>
<td>1.000000</td>
</tr>
<tr>
<td>Interim.</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0.000000</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>1.000000</td>
</tr>
</tbody>
</table>

Total Error = 0.772727

Table 5. Testing set 2 with 4 input variables

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Interim.</th>
<th>High</th>
<th>Total</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>1.000000</td>
</tr>
<tr>
<td>Interim.</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>1.000000</td>
</tr>
<tr>
<td>High</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

Total Error = 0.6363636

Table 6. Testing set 3 with 14 input variables

When examining the three classification tables that correspond to the three adjusted quadratic discriminant models, it is evident that the assignment for new observations is extremely deficient in all three models. This may suggests that the data structure is quite complex and with quadratic polynomials it is not possible to guarantee the appropriate separation of the three drought resistance classes. Another type of representation must be searched that allows the separation of the three drought resistance classes, which must contemplate nonlinear functions that can guarantee the efficient assignment of new observations.

### 3.6 Stage 6. Neural Models Construction

#### 3.6.1 Phase 6.1 Neural Networks Training

For the neural training [1, 3, 5] it was considered the same 3 stratified samples that were used for adjusting the discriminant models of the previous stage. For each sample it was generated a neural model, using the backpropagation learning algorithm in the Statistica Neural Networks ® computational tool [15] and a probabilistic neural network [9] model was constructed using Matlab 6.0® [14].

With the artificial neural networks models obtained using the backpropagation algorithm and probabilistic networks, both gave better observations classification with respect to the discriminant models previously obtained. In the training phase, 5 of 6 adjusted models obtain a perfect classification.
3.6.2 Phase 6.2 Neural networks model or models

Evaluation

In the validation stage, when new observations of the three drought resistance classes have to be assigned to a particular class, the neural networks models that presented a perfect training failed, which can be attributed to overtraining or memorization phenomena. The only model that did not assign all the observations of the adjustment set in the three drought resistance classes suitably, in the validation stage managed to suitably assign 20 of the 22 observations that conformed the testing set. So this model was proposed for new observations classification.

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Interm.</th>
<th>High</th>
<th>Total</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0.000000</td>
</tr>
<tr>
<td>Interm.</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0.000000</td>
</tr>
<tr>
<td>High</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>8</td>
<td>0.250000</td>
</tr>
</tbody>
</table>

Total Error = 0.0909091

Table 7. Classification error using Neural Networks

3.7 Stage 7 Final Results and Conclusions

It was seen, in stage 5, that the discriminant analysis was not efficient in the resistance classes separation, and in the evaluation phase, the assignment of new observations was really deficient. It was expected that the model, in which were included as variable independent or explanatory the 14 points of the induction fluorescence curve, presented an improved classification with respect to the models that only had 4 independent or explanatory variables. That deficient classification is attributed to the complex structure presented by the data, that makes impossible the representation using a quadratic polynomial. Therefore, when using another classification technique as artificial neural networks, the results concerning classification and new observations assignment are satisfactory. The neural networks could capture most of the data complexity, because they use nonlinear functions.

The suggested classification model to be used for the new observations assignment to any of the three drought resistance classes, is the obtained neural model with the stratified sample 1, and backpropagation learning algorithm. This model in addition to being simple from the topology point of view, is the one that maintains better relation between the obtained results in the adjusting and testing phases. The architecture of this model is the following one:

- Artificial neural network with four (4) inputs that correspond to minimum, maximum and intermediate values of the curve (units, tens, hundreds and thousands mW/ms) obtained from the fluorescence signals with linear activation function. One hidden layer with seven (7) neurons and sigmoid activation function. Three (3) neurons in the output layer that correspond to the three drought resistance classes, with sigmoid activation function.

4 Conclusions

The methodological approach given for a patterns recognition general system, was applied in the classification of Pisum sativum plants according to the drought resistance within three classes (high, intermediate and low resistance).

When applying the methodology to the Pisum sativum classification problem, it was observed by means of the exploratory data analysis, that in the 20 values that are obtained from the fluorescence curve, some of them produce the same information, which can generate multicolinearity problems. Thus, of the 20 values that were measured from the curve, 6 are eliminated and a final set of 14 values in mW/ms is obtained.

Although 6 values of the fluorescence curve were eliminated, it was made a principal components analysis to reduce even more the dimensionality of the observations and to select a values set that is representative of the fluorescence curve. Thus, it was obtained a set of 4 values of the curve formed by the minimum, maximum and intermediate values of the curve (units, tens, hundreds and thousands mW/ms).

Three discriminant quadratic models were adjusted, since the linear model is not adapted by the violation of the assumption concerning equality of covariance matrices between the three classes of drought resistance. The results for the classification of new observations are quite deficient, and it can be attributed to the data structure that is complex and with a quadratic polynomial is not obtained a suitable representation for the observations.

The artificial neural networks models adjusted using the backpropagation algorithm and probabilistic networks presented a better representation of the observations with respect to the discriminant models. The model that was proposed for new observations classification is an Artificial neural network with four (4) inputs that correspond to minimum, maximum and intermediate values of the curve (units, tens,
hundreds and thousands) obtained from the fluorescence signals with linear activation function. One hidden layer with seven (7) neurons and sigmoid activation function. Three (3) neurons in the output layer that correspond to the three drought resistance classes, with sigmoid activation function.

In this work is important to emphasize the complement between the statistical techniques and the intelligent techniques like the artificial neural networks. The statistical techniques used in the exploratory analysis are a fundamental tool in the creation of variables sets and observations for the model adjustment in the neural models and in the discriminant models.

In future works could be used, as inputs to the classification models, the obtained principal components. This with the intention to verify if this entails to some improvement in the results. Also, it would be possible to be proposed other schemes for classification using intelligent systems that involve the use of statistical analysis techniques.

5 References