A Methodological Approach for Pattern Recognition System using discriminant analysis and artificial neural networks

Anna Pérez-Méndez*, Elizabeth Torres-Rivas*, Francklin Rivas-Echeverría, Ronald Maldonado-Rodríguez^**

Universidad de Los Andes, Mérida, Venezuela 5101
*Escuela de Estadística
**Laboratorio de Sistemas Inteligentes
^ Bioenergetic Laboratory. University of Geneva, Switzerland

Abstract:- In this work it is presented a methodology for the development of a pattern recognition system using classification methods as discriminant analysis and artificial neural networks. In this methodology, the statistical analysis is contemplated, with the purpose of retaining the observations and the important characteristics that can produce an appropriate classification, and allows, as well, to detect outliers' observations, multicolinearity between variables, among other things.

Key Words:- Classification, Pattern recognition, Discriminant analysis, Artificial Neural Networks.

1 Introduction

In the atmosphere that surrounds to us, much of the information that is handled is presented in form of complex patterns: industrial pieces, faces, written texts, diseases, music, flowers, among others. The study of the mechanisms used by the external signals for stimulating the sensorial organs (seeing, ear, smell, taste and tact) and producing an answer that consist of the labeled, classification or differentiation of such stimuli, has been one of the areas of practically all the disciplines that conform the cognitive science [13, 14, 15].

The human mind classification capabilities, continues being a mostly unknown process, and have not been found conclusive models concerning the nervous system recognition procedure. However, it is admitted that the classification must be made following a general scheme which suggests that before the recognition, a pattern (a form) must be perceived by the sensorial organs; in addition, that pattern or a similar one must have been perceived and remembered previously. Finally, some correspondence must be given between the present perception and the memory [13,14].

That ability that the human being posses, of being able to classify and to recognize forms, has increased the necessity to accumulate them, developing intelligent systems that can help in the solution of complex problems. The study of the human beings capabilities of classification and differentiation, has given origin to a well-known relatively independent discipline like Patterns Recognition.

The patterns recognition is a scientific discipline that studies the mechanisms used by the human mind in order to label complex characteristics and then assigns a name to them. The intention of patterns recognition is the classification and optimal allocation of new forms, or patterns assignment to previously known classes, or categories assignment to a variable. The pattern recognition uses other knowledge areas for discovering the models and tools that allow fulfilling the appropriate assignment [1, 13]. At the moment several areas or approaches in the pattern recognition can be distinguished, using different visions and different tools for trying to solve the classification universal problem. Some of these approaches are:

- Statistical approach
- Numerical approach
- Syntactic approach
- Logical-Combinatory approach
- Neural Networks approach

The technology improvement and the intensive use of computers have impelled the study of pattern recognition, and it has also impelled the creation of patterns recognition systems, that facilitate the accomplishment of multiple tasks in diverse scientific disciplines. For example, in plants taxonomy, patterns recognition is used when it is desired to classify a new specimen within one of the well-known different flowers species [6, 7, 8].

At the moment, there are numerous applications that are being developed in the area of patterns recognition; then it is important the methodologies development for the construction of these systems.

In this work it is presented a methodology as the diverse approaches and techniques fusion result, involved and associated to the patterns recognition [9,10,11,12]. It includes statistical data analysis, with the purpose of retaining the most important observations and characteristics that can help to a better classification.

One of the main objectives of this investigation is to propose a methodological framework for developing a general pattern recognition and classification system using discriminant analysis and artificial neural networks.

The present work is organized as follows: in section 2 fundamental aspects of the discriminant analysis appear, section 3 makes an introduction to artificial neural networks, in section 4 the methodology for the development of pattern recognition systems appears and, section 6 presents the corresponding conclusions, recommendations and further works.

2 Discriminant Analysis

In diverse scientific applications, a common multivariate problem is when a set of observations must be assigned in appropriate form to one of several well-known populations [2, 6, 7].
The discrimination and classification is a multivariate technique whose objective is the separation of different sets based on observations or elements and, the allocation of new observations in some previously defined groups. The discriminant analysis is a separation procedure that often is used on a single sample basis, for investigating the observed differences when the causal relation is not well understood. As classification procedure is less exploratory, since the obtaining of efficient rules of classification is searched that can be used in the allocation of new elements. The discriminant analysis was introduced by Sir Ronald Fisher [6, 7].

The main objectives of discrimination and classification are the following [6]:

- "To describe graphically (in three or less dimensions), or algebraically the object's or observation's different characteristics from several populations or well-known groups. It is desired to find those numerical values that separate or discriminant better the groups ".
- "To order objects or observations in two or more well-known classes. Emphasis is taken in obtaining a rule that can be used for optimally assigning a new individual to a previously determined class ".

3 Artificial Neural Networks

Artificial Neural Networks constitutes a branch of the Artificial Intelligence which tries to emulate the operation and capabilities of biological neural networks related to learning and information processing [1, 5]. Its development has represented a favorable impact for the computer science and its technological applications, and also for other areas like physiology and neurology, with which a very profitable interrelation has been created for the use of cerebral operation models and for the processes interpretation associated with the learning capacities.
Artificial Neural Networks are based on biological behavior, and scientists are concerned in the brain organization, when they consider algorithms and configurations, but the total knowledge about the operation of the brain is so limited. For this reason, Artificial Neural Networks designers should consider the current biological knowledge, looking for structures that execute useful functions. In many cases, this assumption discards biological possibilities or produce networks models that are organically impossible to find or that require a great amount of suppositions about the anatomy and operation of the brain [1].
Due to the advance in computer science, Artificial Neural Networks can be trained using big lots of data, improve their performance and self organization capabilities, fault tolerance and real time operation. This have made that the field of application of Artificial Neural Networks has increased, being used in different tasks as: systems modeling and identification, simulation, processes control, prediction, fault handling, patterns recognition, medical diagnosis, virtual sensors design, etc. [1, 3, 8, 9, 10, 11, 12]

4 An approach for Pattern Recognition and Classification General System Methodology Using Discriminant Analysis and Artificial Neural Networks

In order to be able to make pattern recognition tasks using discriminant analysis and neural networks, it is important to establish a methodological framework, where the diverse aspects to be considered in these tasks are contemplated.

Next, a methodology for developing a Pattern Recognition and Classification System Using Discriminant Analysis Artificial and Neural Networks is proposed.

This methodology is the result of approaches and techniques fusion involved and associated to the patterns recognition [10, 13, 14, 15]. This methodology includes, among other things, statistical data analysis with the purpose of retaining representative observations and variables and for discarding those that are redundant or outliers. This methodology is divided in stages and phases, which are described next.

4.1 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. All the sub processes and variables that take part in the system, and their influence with the problem that is wanted to be solved, are studied. It is essential to study the classification variables, specifying their basic characteristics and their relation with other variables of the system.

4.2 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. In this stage it is verified that the considered system or process, fulfills the considerations necessary for constructing a discriminant model and to train a feedforward neural network; this is, if there exists a representative data set of the problem to be solved.

4.3 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 studied in detail. The available matrix of data is analyzed by means of statistical techniques oriented to the detection of atypical observations, variables transformations if it is required, study the relations between variables, obtain descriptive statistics, among others. The possibility of applying multivariate techniques is evaluated oriented to the reduction of observations and variables.

This stage has three important phases:

4.3.1 Phase 3.1 Matrix of Data Description

This phase consists of the matrix of data description that is going to be used for the classification task. It is necessary to specify the dimensions, the meaning of each variable and the labels used to identify modes or categories. In this phase the missing values are detected and quantified.

4.3.2 Phase 3.2 Software Requirements

In this phase the detailed specification of the software requirements is made. That is, the computational tools that allows to manipulate the information statistically and to fit the discriminant models and also, the evaluation of computational systems that can be used for the neural training.

4.3.3 Phase 3.3 Exploratory Data Analysis

Exploratory Data Analysis is one of the fundamental tools of descriptive statistic [4], whose main mission is to characterize a set of observations that represent the measurement of a variable by means of a reduced number of numerical descriptive measures, with no need to present all the data distribution. These numerical descriptive measures are a good way to summarize the basic characteristics of a variable. These measures form four great groups: central tendency measures, dispersion measures, form measures and non central position measures.

The exploratory data analysis emphasizes the use of graphs or other visual representations to determine the behavior of the data, as well as the possible structures that it present. The use of simple mathematical transformations is verified, that are useful when some assumption is not fulfilled, including the normality and symmetry promotion.

In exploratory analysis, it is very important the measuring of linear association between variables, and it is done examining the correlation matrix. The purpose is to detect which variables are linearly related (directly or inversely) and to detect multicolinearity problems.

Another aspect that must be included in this stage is the "Detection of Atypical Observations". In general the atypical observations or extreme data (outliers) are defined as sample data that seem not to be coherent with most of the data set [4]. The study of atypical observations detection includes numerous techniques but the final decision on the atypical condition or not for a data is based on the investigator criteria, who must take into account if it has happened or not registry errors in data entrance.

If the matrix of data contains great amount of variables, a principal components analysis (PCA) can be made. PCA is one of the most spread multivariate techniques, and which main target is to reduce the original data set dimensionality, generating new variables set (linear combinations of the original variables) called "principal components" which are uncorrelated and are independent. These new variables or principal components, are ordered in such a way, that the first components retain most of the presented variation in all the original variables [6, 7].

4.4 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:

4.4.1 Phase 4.1 Data processing

The processing of the variables selected as independent or explanatory variables for the discriminant and neural models is made. The treatment of atypical observations is made; mathematical transformations are applied to the variables that require it to promote normality or symmetry.

4.4.2 Phase 4.2 Adjustment Set Selection

In order to construct the discriminant and neural models, it is necessary to make a partition of the original set of data in two sets: a set used for constructing the models that will be denominated "adjustment set", and the other set used for evaluating or for testing these models. The "test set" is the complement of the adjustment set. For the selection of the observations that will constitute the "adjustment set" and "test set" for the discriminant and neural models, some procedure of random sampling can be used.

4.5 Stage 5. Discriminant Analysis

It is made the adjustment and evaluation of discriminant models. This stage contemplates two important phases:

4.5.1 Phase 5.1 Discriminant Models Construction

The construction of linear or quadratic discriminant models is made, considering the variables and observations selected in the previous stage. Initially it is verified if observations corresponding to each class are linearly separable (the simplest case) and if all the assumptions that the linear discriminant analysis implies, for example normality and variance and covariance matrices equality are fulfilled. If it is not linearly separable, the construction of more complex models, like for example the quadratic discriminant analysis is given. The classification results in each case should be evaluated and then it can be suggested the most appropriate model.

4.5.2 Phase 5.2 Discriminant Models Evaluation

It is evaluated the discriminant models obtained in the previous phase, considering the observations and variables that conform the validation set. The objective is to evaluate the classification or assignment of new observations to the previously defined classes and to obtain a classification error measurement. The best model is obtained from comparing the classification results obtained with the linear and the quadratic model.

4.6 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. In this stage it is evaluated the networks types that can better represent the original data, and finally the evaluation or test of the obtained models is made. Like in stage 5, this stage has two important phases:

4.6.1 Phase 6.1 Neural Networks Training

In this phase it is chosen the neural network topology, (hidden layers, number of neurons in the hidden layers, activation functions, etc.). The network inputs were selected in stage 4 and the network's outputs corresponds to the classification variable classes or categories. In this phase, different types of neural networks can be evaluated.

4.6.2 Phase 6.2 Neural networks model or models Evaluation

In this phase it is evaluated the obtained neural models in the previous phase, considering the observations and variables that conform the testing set. The objective is to evaluate the generalization capabilities of the neural network obtained, given in terms of the classification or assignment of new observations to the classes previously defined and to have a classification error measurement for comparing the models until finding the best one.

4.7 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.

4.8 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.

In figure 1, it is presented the methodology stages scheme for developing a patterns recognition system.

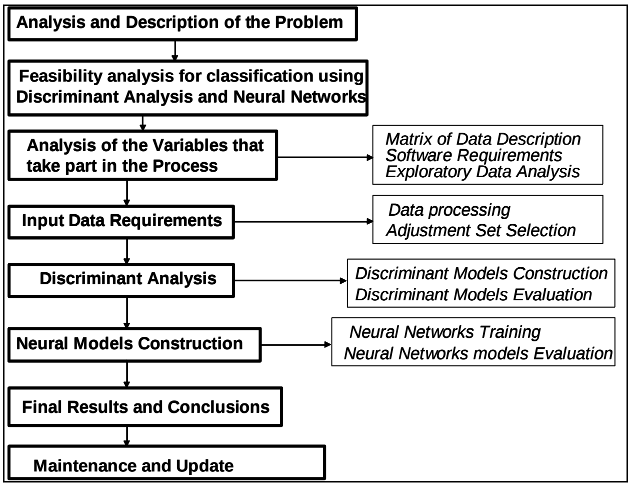
Analysis and Description of the Problem
Feasibility analysis for classification using Discriminant Analysis and Neural Networks
Analysis of the Variables that take part in the Process
Matrix of Data Description
Software Requirements
Exploratory Data Analysis
Input Data Requirements
Data processing
Adjustment Set Selection
Discriminant Analysis
Discriminant Models Construction
Discriminant Models Evaluation
Neural Models Construction
Neural Networks Training
Neural Networks models Evaluation
Final Results and Conclusions
Maintenance and Update

Figure 1. Methodology General Scheme for developing a pattern recognition system

5 Conclusions

In this work appears a methodology for the development of a pattern recognition general system, using discriminant analysis and neural networks as classification procedures.

This methodology is the result of the fusion of approaches and techniques associated to the universal problem of classification. It includes, statistical analysis of data, mainly statistical descriptive, with the purpose of retaining the representative observations and variables or characteristics; and to discard the information that is redundant or atypical.

This methodology is being applied in the classification of *Pisum sativum* plants according to the drought resistance within three classes (high resistance, intermediate resistance and low resistance) [16].

6 References

[1] Aguilar Castro Jose y Rivas Echeverría Francklin. Introducción a las Técnicas de Computación Inteligente. Editorial Meritec. Mérida Venezuela 2001.

[2] Anderson T. W. An Introduction to Multivariate Statistical Analysis. Wiley Series in Probability and Statistics. Tercera Edición. 2003

[3] Colina Morles Eliezer y Rivas Echeverría Francklin. Inteligencia artificial Aplicada. Universidad de Los Andes. Postgrado en Ingeniería de Control y Automatización. Mérida Venezuela 1998.

[4] Freixa Monserrat, Salafranca Luis, Ferrer Ramón, guardia Joan y Turbany Jaime. Análisis Exploratorio de Datos. Promociones y Publicaciones Universitarias. Primera Edición. Barcelona España 1992.

[5] Hagan Martin, Demut Howard and Beale Mark. Neural Network Design. An International Thompson Publishing Company. 1996

[6] Jonson Richard and Wichern Dean. Applied Statistical Analysis. Cuarta Edición. Prentice Halll 1998.

[7] Morrison Donald F. Multivariate Statistical Methods. Tercera Edición. Mc Graw Hill Publishing Company 1990.

[8] Maldonado Rodríguez Ronald, Pavolv Stancho, Gonzálesz Alberto Oukarroum Abdallah, strasser Reto J. Can Machines Recognise Stress in Plants? © Springer-Verlag 2005

[9] Darcy Novoa, Anna Pérez, Francklin Rivas. Fault Detection scheme using Neo-fuzzy Neurons. IASTED International Conference on Intelligent Systems and Control Agosto 2000.Honolulu, Hawaii. USA

[10] Pérez M. Anna Gabriela, Nava Puente Luis, Rivas Echeverría Francklin. Metodología para el Desarrollo de Sensores Virtuales Basados en Redes Neuronales. Universidad de Los Andes. Facultad de Ciencias Económicas y sociales. Escuela de Estadística. Mérida – Venezuela 2001

[11] Rivas-Echeverría, F., Cerrada M. And Aguilar, D, “Fault detection scheme using neural networks with fuzzy preprocessing” in Proc. 14th World Congress of IFAC 1999.

[12] Rivas–Echeverría F., Olivares M., Pensa R. “Probabilistic Neural Network Based System for Well Characterization in Oil Industry”. Advances in Neural Network World. WSEAS Press. 2002

[13] http://www-etsi2.ugr.es/depar/ccia/rf/www/tema1_00-01_www/node2.html

[14] http://ccrma.stanford.edu/~juanig/articles/charlAndes/Elementos_Sistemas_Reconoci.html

[15] http://www.geocities.com/CapeCanaveral/Hangar/4434/pattern.html

[16] Pérez Anna, Torres Elizabeth, Maldonado, Ronald, Rivas Francklin. “Pisum Sativum Classification using discriminant analysis and Neural Networks”. 6th WSEAS International Conference on Neural Networks. Lisbon, Portugal 2005.