Title: Learning a rule-based model to describe weed infestations in terms of soil factors

Authors:

A. Ribeiro, B. Díaz and M.C. García-Alegre
Industrial Automation Institute. Spanish Council for Scientific Research.
28500 Arganda del Rey. Madrid. Spain.
Tel. 91 871 19 00
{angela, bdiaz, maria}@iai.csic.es

D. Ruiz, J. Barrosos, C. Fernández-Quintanilla
Centro de Ciencias Medioambientales (CCMA)- Spanish Council for Scientific Research.
Serrano 115 B, 28006 Madrid. Spain
Tel. 91 562 50 20

Abstract:

Data mining is justified in those applications where conventional analysis methods are not able to extract useful, task-oriented knowledge. Site specific treatment of weeds in Precision Agriculture, is one of those application where it is essential to know which factors determine a high occurrence of weeds, due to environmental and economic costs associated consequences. Some statistical studies carried out previously, using regression techniques searching for relations between individual soil variables and wild oat occurrence, have concluded that there is not a clear linear relationship among the analysed variables. However, farmers observe that wild oat grows better in specific locations, where probably the physical and chemical conditions are more suitable. Present work proposes a fuzzy rule based model, generated from a supervised learning process, to describe the complex relationships among weed occurrence and some soil properties. A genetic algorithm is used to derive Mandami-type fuzzy rules by driving a search in the space of possible solutions or models. This linguistic rule model has the advantage to be both intuitively and directly interpretable by the human expert involved in the field tasks.

Keywords: Supervised Learning, Weed Control, Fuzzy Rule Based Systems, Genetic Algorithms.

Topics:

- Aprendizaje Automático, Descubrimiento de Conocimiento y Minería de Datos.
- Computación Evolutiva, Algoritmos Genéticos y Redes Neuronales

Section: Paper track. The article never has been sent to other conference.

Learning a rule-based model to describe weed infestations in terms of soil factors

A. Ribeiro, B. Díaz and M.C. García-Alegre
Industrial Automation Institute. Spanish Council for Scientific Research.
28500 Arganda del Rey. Madrid. Spain.

D. Ruiz, J. Barrosos, C. Fernández-Quintanilla
CCMA- Spanish Council for Scientific Research.
Serrano 115 B, 28006 Madrid. Spain

Abstract
Data mining is justified in those applications where conventional analysis methods are not able to extract useful, task-oriented knowledge. Site specific treatment of weeds in Precision Agriculture, is one of those application where it is essential to know which factors determine a high occurrence of weeds, due to environmental and economic costs associated consequences.Some statistical studies carried out previously, using regression techniques searching for relations between individual soil variables and wild oat occurrence, have concluded that there is not a clear linear relationship among the analysed variables. However, farmers observe that wild oat grows better in specific locations, where probably the physical and chemical conditions are more suitable. Present work proposes a fuzzy rule based model, generated from a supervised learning process, to describe the complex relationships among weed occurrence and some soil properties. A genetic algorithm is used to derive Mandami-type fuzzy rules by driving a search in the space of possible solutions or models. This linguistic rule model has the advantage to be both intuitively and directly interpretable by the human expert involved in the field tasks.

1 Introduction

Precision Agriculture (PA) searches for techniques that allow to minimise the use of agrochemical products whilst ensuring that weeds, diseases and pests are effectively controlled and crops are provided with adequate nutrients [1]. The selective spraying of weed in cereal crops requires the elaboration of risk maps to derive location and concentration of the products to be applied. Consequently, precise maps are essential for an accurate PA practice. However, the risk map generation is a complex task, due to the great number of variables of different nature that are involved in the weed evolution. On the other hand, it is an observable fact that in a crop field, weeds tend to appear in aggregated patterns while some regions remain weed-free, but the causes of these spatial variations are not yet clearly ascertained. Some studies has been conducted to find relations among soil properties and weed occurrences [2], [3] deriving only relations among the presence of certain species under specific soil conditions. Moreover, such data were previously analysed by statistical classical methods and no consistent correlation was found. In fact the contradictory results suggest that soil properties by themselves are not sufficient to explain wild oat distribution in the fields. Therefore it is not possible to achieve an accurate analytical model of the weed evolution only from soil properties. As a consequence, to model weed evolution, more variables not always available, must be

taken into account such as, the field history, the landscape characteristics, the weather or the seed dispersion.

This search for data relationship constitutes a perfect test-bed to investigate inferences within the framework of a data mining approach. The objective of this research is to find out complex relationships among the soil parameters and the abundance of winter wild oat (Avena sterilis L.) in dry-land cereal fields. This problem can be focused as the problem of obtaining an adequate description of the concept “a high wild oat amount” versus the concept "a NOT high wild oat amount"; in other words a descriptive model that covers two sets of well-classified input data that represent both concepts. A descriptive Fuzzy Rule-Based System (FRBS) has been selected to model previous concepts since this kind of descriptive approach conveniently frames the system behaviour with a set of linguistic expressions in natural language, being the model interpretability its primary aim [4]. Consequently, the derived model is composed of a set of descriptions directly interpretable by a human expert

A Genetic Algorithm (GA) based search is used to derive the most suitable model (descriptive FBRS), since it offers a powerful search methodology, domain independent. In fact the GAs have been used in machine learning process [5] to obtain production systems from an input set of examples (supervised learning). Moreover, GAs have proven to be a powerful tool for automatic FRBS definition, since adaptive control, learning, and self-organisative FRBSs can be considered in many cases as optimisation or search processes [6].

Detailed aspects and reasons that have motivated both the current problem solving approach and the proposed model can be found in the following paragraphs, organised as follows. Section 2 presents the problem and the main characteristics of the data. Section 3 provides a brief overview of fuzzy rule-base systems and describes current approach. Section 4 outlines the more illustrative aspects of the genetics algorithms and describes present approach to discover the descriptive model. Section 5 reports the computational results. Finally, in section 6 the advantages of the proposed approach are discussed as well as the future research activities.

2 The Data Sets

Soil sampling by grid is the most adopted technique [7],[8],[9], as grid methods appropriately display spatial and temporal variability in soil attributes for an efficient action/treatment. Thus, data have been obtained from a quadrangular grid sampling carried out in four barley fields and in two different locations in SE Madrid. Field sizes range from 0.5 ha. to 1.6 ha. In each grid point, soil samples and wild oat abundance data, were gathered. The soil parameters analysed were pH, organic matter (OM), nitrogen (N), phosphorus (P), potassium (K), and sand, silt and clay percentages. Wild oat counting was performed using a square of $0.1m^2$, sampling a total of 146 points.

To reduce other factor effects such as the field history and the landscape characteristics in the weed evolution in each field, as well as to accomplish future comparisons with other sampling experiments, the recollected data have been normalised within the range of [0-1]. As a result, the “a high wild oat amount” concept becomes a relative concept.

Now then, supervised concept learning involves inducing concept descriptions from a set of positive and negative examples of the target concept [10]. Examples are represented as points in a n-dimensional feature space which is a priori defined and for which all the legal values of the features are known. Concepts are therefore represented as subsets of points in the given n-dimensional space. The biological data involved in the experiment have been divided in two sets, positive and negative examples, using the quantity of wild oat observed at each point. The wild out seed distribution for both fields is displayed in figure 1; the map was obtained from an interpolation krigging process and adequately visualises the weed distribution. From the representation displayed in figure 1, a good threshold to build the training sets can be derived, 0.2. This threshold defines as a positive example the tuple (pH, OM, N, P, K, sand, silt, clay) when wild oat seed amount is higher than 20% of the largest seed value in the field, and as a negative example otherwise. Consequently, the positive training set has 39.7% of the input data and the negative training set the rest of the departure examples.

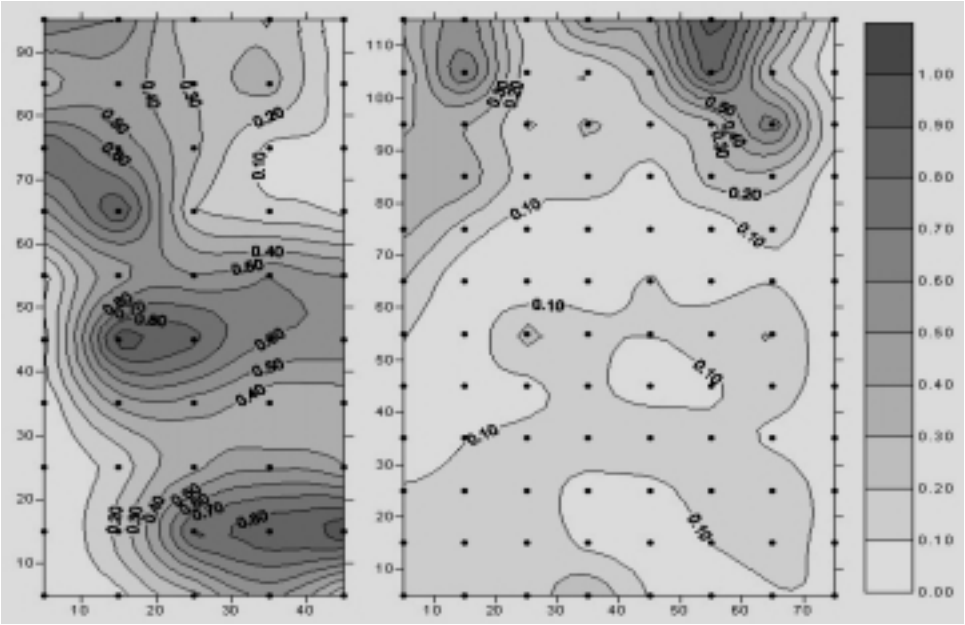

Fig. 1. The wild oat distribution in the selected fields.

3 Fuzzy Rule-Based Systems and Knowledge Representation

Nowadays one of the most important areas in Fuzzy Set Theory and Fuzzy Logic [11] are the Fuzzy Rule-Based Systems (FRBSs) which constitute an extension of the classical Rule-Based Systems. These systems use IF-THEN rules, where both antecedent and consequent are composed of fuzzy propositions instead of classical logic propositions. FRBSs have been successfully applied to a wide range of real-

world problems in different areas [12], [13], [14], [15], [16], [17]. They provide a good platform to deal with noisy, imprecise or incomplete information, which is often handled in many human-cognition tasks. To model a real systems by means of FRBS has many advantages such as; the fuzzy rules have always a interpretation clearer that classical analytic approaches, since each rule defines the function that describes the system behaviour in a small subspace of the working space. In other words, the global model is composed of a set of local models, each one driven by a rule, which facilitates the model interpretation.

Present work follows Mamdami FRBSs guidelines that proposes a linguistic expression with a fuzzy variable in the consequent of each rule of the Fuzzy Rule Base (FRB); thus the FRBs is composed of a collection of fuzzy rules with the following structure:

$$R_i : IF\ x_{i1}\ is\ A_{i1}\ and \ldots and\ x_{in}\ is\ A_{in}\ THEN\ y_i\ is\ B_{in}$$

where $x_{i1}\ldots x_{in}$ input variables and y_i is the output variable.
On the other hand, a Mamdami type system is a descriptive approach and $x_1\ldots x_n$ and y_i are real world variables values. Each variable is described in terms of several fuzzy sets values $\{A_{ij}$ or $B_{il}\}$and each A_{ij} or B_{il} fuzzy set is described by means of a membership function, being this mapping is uniform for all rules in the FRB. Moreover the Mamdami Fuzzy Rule-Based System gives an useful frame to describe expert knowledge by a set of linguistic rules and allows the combination of rules derived from the data that reflect the system knowledge. A FBRSs is composed of four modules: *Fuzzy Knowledge Base*, *Fuzzyfication Interface*, *Fuzzy Inference System* and *Defuzzyfication Interface*

The *Fuzzy Knowledge Base* is the essential part of a descriptive FRBS and is composed of the Fuzzy Rule Base (FRB) and the DataBase (DB). The FRB packs a set of linguistic rules with fuzzy variables; rules are linked so as all of them can be fired simultaneously for a specific input. On the other hand, the DB holds the definition of the fuzzy set associated to the linguistic terms defining each variable value in the FRB. The *Fuzzyfication Interface* allows the FRBS translate the crisp inputs to linguistic labels and the opposite to get crisp outputs, establishing a correspondence between crisp input values and the defined fuzzy sets. The *Fuzzy Inference System* is based in the Generalisation of the Modus Ponens, which is an extension of the Classical Logic Modus Ponens. Once the fuzzy inference has been applied for each of the *m* rules in the FRB, *m* fuzzy set outputs are obtained. These outputs represent the fuzzy actions inferred by the FRBS from the input data. Finally the *Defuzzification Interface* is an algorithmic process to aggregate the *m* actions to obtain a crisp value.

In presented approach, the fuzzy sets corresponding to each linguistic label {Low, Medium, High} are represented by means of trapezoidal membership functions, and their shape and slope are fixed from the expert criterion (see figure 2) while a supervised learning process is conducted to derive the FRB. Rules in the Fuzzy Rule Base are, for example, as follow:

IF K is High and Sand is Low and OM is Medium THEN Wild_Oat is High

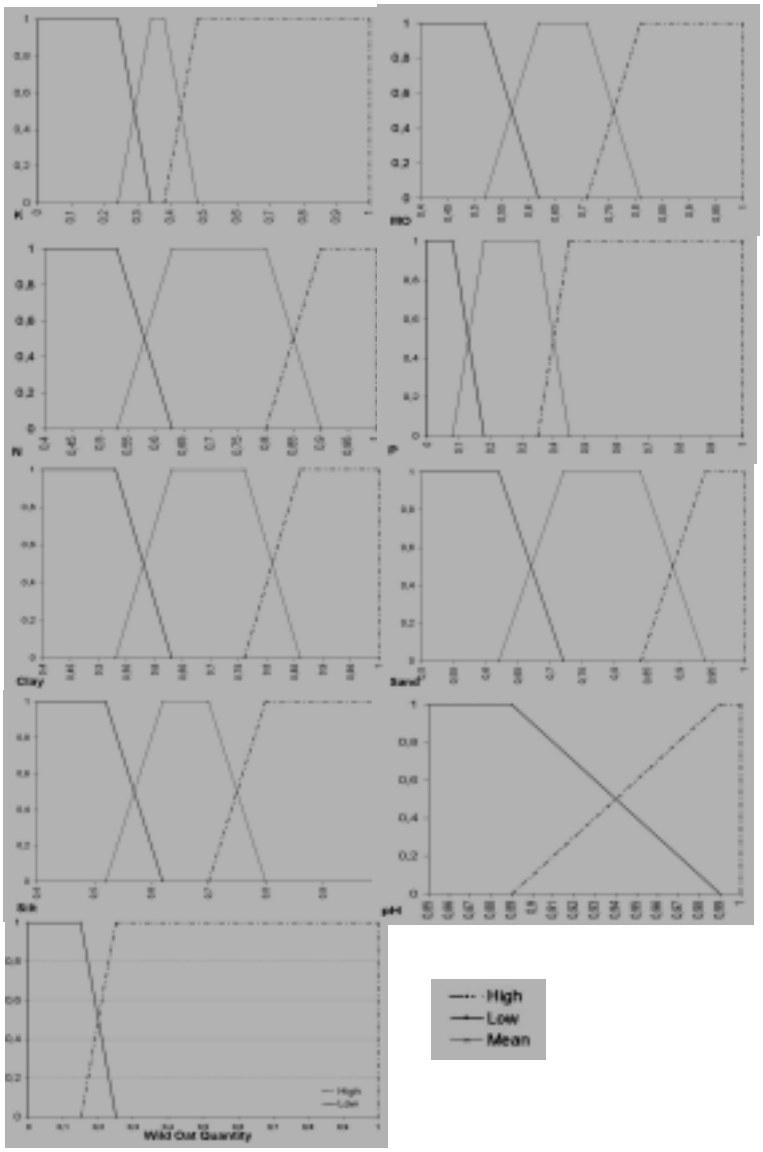

Fig. 2. Membership functions for the sampled variables.

Finally, all algorithmic processes, the fuzzyfication, the fuzzy inference and the defuzzyfication are performed with the aid of the FuzzyShell[1] programming environment [18]. The aggregation operator used in the defuzzyfication step is based

[1] FUZZYSHELL is a CSIC registered trademark 1643983.

on a centre of mass algorithm (CO) that weights the output of each rule in the FRB with the truth degree of its antecedent.

4 Genetic Algorithm and Search Process

Genetic algorithms (GAs) [19] are search and optimisation techniques based on a formalisation of natural genetic processes. Over the last few years, the advantages of the GAs have extended its use in the development of a wide range of approaches for designing FRBSs, namely Genetic Fuzzy Rule-Based Systems (GFRBSs). These Systems are an important branch of the Soft Computing area [20] as demonstrated by the large number of contributions published during the last decade.

The basic idea underlying genetic process is to start with a population of randomly generated solutions, namely chromosomes, that define the first generation *(G(0))*, from which evolution begins. While a specific ending condition is not met, each chromosome is evaluated with respect to its ability to solve the target problem. This evaluation is performed by means of a fitness function. Then a new population is created *(G(t+1))*, by applying a set of genetic operators to the individuals of the generation *G(t)*. The more common operators are the selection, the crossover and the mutation.

In order to apply a GAs to a particular problem, we need to select an internal representation of the search space and define a fitness function, which assigns an utility level to each candidate solution.

4.1 Representation in the Search Space

The traditional internal representation used by GAs involves fixed-length (generally binary) strings to represent point in the space to be searched. In the current application, three linguistic labels are defined for each physical variable {low, medium and high}, except for the pH and Wild_Oat variables, which have only two linguistic labels {low, high}. The definition of an uniform structure facilitates both codification and decodification of the chromosomes. For this reason two bits are used to code each label, so that codification is as follow: (low, 01), (medium, 10), (high, 11) and, for pH and Wild_Oat variables, (low, 10) and (high, 11). The configuration 00 for the three-label variables, and 01 and 00 for the two-label ones are used to represent the absence of antecedents or consequent respectively. Using this representation the antecedents and the consequent of each rule is internally represented by means of a binary string, such as:

pH	OM	N	P	K	Sand	Silt	Clay	Wild_Oat
-	Medium	-	-	High	Low	-	-	Low
00	10	00	00	11	01	00	00	10

Specifically this binary-string characterises the following rule:

IF OM is Medium and K is High and Sand is Low THEN Wild_Oat is Low

Since the FRBs is composed of one or more fuzzy rules, an specification of how GAs are used to guide the evolution of sets of rules, is needed. There are currently two basic strategies: the Michigan and the Pittsburgh approaches. Systems using the Michigan approach maintain a population of individual rules, which compete with each other for space and priority in the population. Opposite, systems using the Pittsburgh approach maintain a population of rules sets, which compete among themselves. In present work a Pittsburgh approach has been used, thus each individual in the population is a fixed-length string representing an unordered set of rules. Also, for each rule *i*, the binary configuration 00 for every antecedent as well as the first bit of the consequent equal to 0 represents the absence of rule *i* in the FRB. Consequently, the number of rules in a particular individual can range from 0 to 10.

With the selected representation genetic operators are adapted with minimal variations. In fact, a classical two-point crossover and an usual bit-level mutation operator have been used.

4.2 The Fitness Function

In addition to selecting a good representation it is important to define a good payoff function. In our case the selected fitness function [21] evaluates the quality of a chromosome. This means that the FRB codified in the chromosome classifies a number of well-known database records, and the predictive performance of FRB is computed. The fitness function combines two indicators namely the sensitivity (*Se*) and the specificity (*Sp*), defined as follows:

$$Se = tp/(tp + fn)\,. \tag{1}$$

$$Sp = tn/(tn + fp)\,. \tag{2}$$

Where tp, fp, tn and fn are the number of true positives, false positives, true negatives and false negatives, respectively. The true positives and the true negatives are FRBS well-classified examples, while false positives and false negatives are miss-classified examples. The fitness function is defined as the product of the previous indicators as follow:

$$fitness = Se * Sp \tag{3}$$

Finally, the fitness can be completed by a term that enforce the AG search to produce FRB as short as possible. The motivation is that the comprehensibility of the rule is inversely proportional to its size. This term can be defined as follows:

$$S_y = \frac{(maxrules - 0.5 * numrules - 0.5)}{(maxrules - 1)} \tag{4}$$

Where *numrules* is the current number of rules in the chromosome and *maxrules* the maximum number of rules that can be represented in a chromosome. The equation [4] shows its maximum value 1, when the FRB codified in the chromosome has only

a rule and a minimum value of 0.5 when the number of rules contained in the FRB is equal to the maximum allowed. The minimum is selected to penalise large-size individual without forcing them to disappear. The new fitness function could be defined as follows:

$$fitness\_size = S_e * S_p * S_y \quad (5)$$

5 Results

The input data have been split in two sets of data: the training set and the verification set. The training set contains approximately the 55% of the total input data, 40 positive examples and 40 negative examples. A lot of tests have been performed in order to achieve an adequate configuration for the genetic search. The results and some parameters of two experiments, one with fitness value obtained with equation (3) and another with equation (5), are displayed in table 1. The initial population is randomly generated. The 2 points crossover was selected and applied with a 0.5 probability in the two experiments. The bit-by-bit mutation operator was used with a probability equal to 0.01 in experiment 1 and equal to 0.005 in experiment 2. The reproduction mechanism was the roulette wheel selection, but to force the convergence in experiment 1 best chromosome has been preserved inter generations, and an elitist model has been used in the experiment 2. The verification results are shown also in table 1. In this case the FRBs obtained in the two experiments are contrasted against the verification set formed by 58 positive and negative examples. Best classification rate is achieved in experiment 2, in spite of the fitness value being worse than that obtained in experiment 1. This suggests that a better solution can be found with a FRB with more than two rules. As a consequence, more experiments are needed to tune the proposed approach.

Table 1. Mainly results for two experiments.

Experiments	*Rules number/maximum*	*fitness*	*fitness_size*	*Classification task*	
				Successes (%)	Faults (%)
1	7/10	0.734	-	65	34
2	2/30	0.585	0.575	70	30

The best FRB from a classification point of view has the two following rules:

IF N is Medium and MO = Medium and Clay is Medio THEN Wild_Oat is High

IF pH is High and P is Medium and Clay is High THEN Wild_Oat is Low

Finally the application of these simple rules gives rise to an increase in the accuracy of the weed risk maps.

6 Conclusion and Future Research

Present work proposes an approach to learn a descriptive rule-base model from a sample data set to extract relevant information on the complex relationships among some of the variables involved in the weed evolution to help in the construction of more accurate risk maps. With this idea in mind, a learning system that integrates a Fuzzyfication Interface, a Fuzzy Inference System, a Defuzzyfication Interface and an AG to perform the search of the best FRB to appropriately describe the concept "a high wild oat amount", has been designed and developed. Previous works with classical algorithms were not able to clearly demonstrate a correlation among soil factors and weed occurrence.

In the problem here presented, there is not a way of finding out whether or not there exist complex relationships, except by conducting a machine learning/data mining process. FRBSs have been selected due to the expressiveness of the linguistic rules to formulate models directly understandable by the expert. On the other hand the initial results support the fact that GAs can be used as an effective FRB learner.

Future research includes demonstrations with data coming from different fields, tests selection mechanisms, and the definition of a fitness function that incorporates a new term to evaluate the number of antecedents in each rule to reward the shorter length rules.

7 Acknowledgements

The authors wish to thank Domingo Guinea and Lía García-Pérez for their valuable help, the Spanish Science and Technology Commission for funding this research through the Grant AGF1999-1125-C03-03, and the Ministry of Science and Culture for a pre-doctoral grant.

References

1. Kropff, M.J., Wallinga J., Lotz LAP Modelling for precision weed management. In Precision Agriculture: spatial and temporal variability of environmental quality. Wiley, Chchester (1997) 182-204.
2. Dieleman, J. A. et al.,. Weed Science 48: (2000)567-575.
3. Rew, L. J. & R. D. Cousens. Weed Research 41: (2001)1-18.
4. Sugeno M. And Yasukawa T. A fuzzy-logic-based approach to qualitative modeling. IEEE Transaction on Fuzzy Systems 1(1): (1993) 7-31.
5. Grefensette J.J. (Ed) Genetic Algorithms for Machine Learning. Kluwer Academic. (1994).
6. Cordón O. and Herrera F. Hybridizing Genetic Algorithms with Sharing Scheme and Evolution Strategies for Designing Approximate Fuzzy Rule-Based Systems. Fuzzy Set ans Systems 118:1 (2001) 47-64.

7. Colliver, C.T., Maxwell B.D., Tyler D.A., Roberts D.W., and Long, D.S. Georeferencing wild oat infestations in small grains accurancy and efficiency of three weed survey techniques. In Proceedings 3rd International Conference on Precision Agriculture (Ed. P.C. Robert and others) USA (1996) 453-463.
8. Christensen, S. and Heisel T. Patch spraying using historical, manual and real time minitoring of weeds in cereals. Z. PflKrankh. PflSchutz, Sonderh. XVI, (1998) 257-263.
9. Clay, S. A., Lems, G. J., Forcella, F., Ellsbury, M.M., and Carlson, C.G. Sampling weed spatial variability on a fieldwide scale. Weed Science 47, (1999) 674-681.
10. Michalski R. S. A theory and methodology of inductive learning". Machine Learning, an Artificial Intelligence approach, volumen 1. Michalski, R.S., Carbonell J.C., and Mitchell T. M., editors. Morgan Kaufmann, San Mateo, California (1983)
11. Zadeh L. A. Fuzzy sets. Information and Control 8: (1965) 338-353.
12. Hirota K. (Ed). Industrial Applications of Fuzzy Technology. Springer-Verlag, (1993).
13. Garcia-Alegre M.C., Ribeiro A., Gasós J., Salido J., Optimization of fuzzy behavior-based robots navigation in partially known industrial environments. IEEE Inter.Conf.on Industrial Fuzzy Control & Intell.Syst. Houston,TX (1993) 50-54.
14. Bardossy A. And Duckstein L. Fuzzy Rule-Based Modeling with Application to Geophysical Biological and Engineering Systems. CRC Press. (1995).
15. M.C. García-Alegre, D. Guinea, R. Gonzalez-Bueno, A. Ribeiro. "Fuzzy Diagnose Microcontroller Based System for Air Quality Surveillance". 10th IEEE International Conference on Fuzzy Systems, Melbourne. December (2001).
16. García-Pérez L., Marchant J., Hague T. and García-Alegre M.C. Fuzzy Decision System for Threshold Selection to Cluster Cauliflower Plant Blobs from Field Visual Images. SCI2000, Orlando, (2000) 23-28.
17. Ribeiro A., Fresno V., García-Alegre M.C., Guinea D. A Fuzzy System For The Web Page Representation. Intelligent Exploration of the Web, (P.S.Szczepaniak, J.Segovia, J.Kacprzyk, L.A.Zadeh, Eds.), Studies in Fuzziness and Soft Computing (J.Kacprzyk Ed.), Springer-Verlag, 2002. (in press)
18. Gasós J., Fernandéz P.D., García-Alegre M.C., Garcia Rosa R. Environment for the development of fuzzy controllers. Proc. Intern.Conf. on A.I. Applications & N.N. (1990) 121-124,.
19. Michalewicz Z. Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag, (1996).
20. Bonissone P.P., Soft computing: the convergence of emerging technologies. Soft Computing 1:1 (1997) 6-18.
21. Bojarczuk C. C., Lopes H. S., and Freitas A. A. Data Mining with Constrained-Syntax Genetic Programming: Applications in Medical Data Set. Data Analysis in Medicine and Pharmacology (IDAMAP-2001), a Worshop at Medinfo-2001. London, UK (2001).