

# Web Page Classification: A Soft Computing Approach

Angela Ribeiro<sup>1</sup>, Víctor Fresno<sup>2</sup>, María C. Garcia-Alegre<sup>1</sup> and Domingo Guinea<sup>1</sup>

<sup>1</sup>Industrial Automation Institute. Spanish Council for Scientific Research.  
28500 Arganda del Rey. Madrid. Spain.

{angela, maria, domingo}@iai.csic.es

<sup>2</sup>Escuela Superior de Ciencia y Tecnología

Universidad Rey Juan Carlos

v.fresno@escet.urjc.es

**Abstract.** The Internet makes it possible to share and manipulate a vast quantity of information efficiently and effectively, but the rapid and chaotic growth experienced by the Net has generated a poorly organized environment that hinders the sharing and mining of useful data. The need for meaningful web-page classification techniques is therefore becoming an urgent issue. This paper describes a novel approach to web-page classification based on a fuzzy representation of web pages. A doublet representation that associates a weight with each of the most representative words of the web document so as to characterize its relevance in the document. This weight is derived by taking advantage of the characteristics of HTML language. Then a fuzzy-rule-based classifier is generated from a supervised learning process that uses a genetic algorithm to search for the minimum fuzzy-rule set that best covers the training examples. The proposed system has been demonstrated with two significantly different classes of web pages.

## 1 Introduction

The Web has grown by leaps and bounds, making it increasingly difficult to locate useful information. Moreover, Internet users numbered 30% more at the end of 2002 than at the end of 2001 [1]. Under these circumstances the need for tools that adequately extract information from the Web is becoming a major fact.

Classic information retrieval (IR) techniques are normally used to obtain information from the Internet [2], but problems frequently appear when IR techniques are applied. These problems may be due to either the enormous number of pages online or the continuous changes that happen in those pages, but they may also occur because Internet users are significantly different from IR techniques' traditional user groups. In the Web, there is no standard or style rule, and page content is created by a set of heterogeneous, autonomous users. Moreover inherited IR technology has not progressed rapidly enough to take into account the Web's needs. As a consequence, search engines retrieve only a small fraction of the total number of available documents, and only a part of what is retrieved is significant. In fact, when a user places a search request, most known web-search services deliver a ranked list of web pages wherein different topics or aspects of a topic are jumbled together. As stated by

Chen and Dumais [3], users often prefer a search result organized as a category-based view of retrieved documents, to enable them to pick out the most relevant information in the shortest time.

From a general point of view, the first step in tasks such as automatic summarization, text classification, information retrieval, information extraction and text mining for any type of document is to obtain an adequate data structure to represent the text and ease digital processing. However, it is difficult to gain access to text information, since in texts the relationship between structure (usually a sequence of characters) and meaning is not as straightforward as it is in numeric data. Thus, the data structure that represents a web page is essential, since the success of the analysis is strongly dependent on the correctness of the text representation, which selects a document's most relevant aspects. Without a proper set of features, a classifier will not be able to discriminate categories accurately. Text and unstructured documents have traditionally been represented in a vector-space model [4] [5]. The vector representation or bag words takes single words extracted from the training corpus as features. This representation ignores the sequence in which the words occur and is based only on the statistics of single independent words. A feature may be either Boolean or frequency based. Variations on feature selection include removing infrequent words and stop words. In all the descriptions of vector space, the location of the words in the document is lost, and vectors usually have a high dimensionality (of  $10^4$  to  $10^7$  components) that prevents the use of knowledge-extraction algorithms [7]. A quite complete relation of web content mining methods for both unstructured and semi-structured documents is presented in [6].

For hyperlink texts, several techniques have been proposed for classifying a web page. Some are based on web content (the intra-document structure), and others, on the structure of the hyperlinks in the web itself (inter-document structure). Here researchers are inspired by the study of social networks and citation analysis [8]. Furthermore, in order to extract rich, predictive features from both sources some researchers combine hyperlinks and textual information. The major point is to be aware that in real-world cases an inspection out of the document can disturb classification performance. Even when there are no hyperlink regularities, no benefit can be expected from using hyperlinks, and in this case the best course is to use standard text classifiers on the text of the document itself. In [9] a study is shown of approaches for hypertext categorization exploring different hypotheses about the structure of the hypertext. This paper's approach was motivated by relevant conclusions of study [9], such as: a) The identification of hypertext regularities in the data and the selection of appropriate representations for hypertext are crucial for an optimal design of a classification system; b) The recognition of useful HTML fields in hypertext pages to be jointly considered with the text contained in the web page improves classification performance.

This paper proposes a supervised learning process based on a genetic algorithm to obtain a classifier. The classifier is expressed in terms of a fuzzy-knowledge-based system, and it is obtained from input web-page examples expressed as feature vectors. The vector contains the web page's most representative words, associated with a value that characterizes each word's degree of relevance in the hyperlink text.

The rest of the paper is organized as follows: Section 2 presents some basic ideas of the web-page representation technique proposed herein [10]. Section 3 is devoted

to the description of the learning process that generates the fuzzy classifier. Section 4 presents some classifier performance data, and lastly, Section 5 outlines some conclusions.

## 2 A Fuzzy Approach to Web-Page Representation

A two-dimensional vector, namely a feature vector composed of features with an associated weight, has been used to represent a web page. Features are words extracted from the page content, and weights are numeric values that evaluate the appropriateness of the word to represent the web page's text. The word-relevance estimate takes into account a combination of traditional text-representation methods and some specific characteristics of HTML language. Web documents are built as a combination of HTML tags and text information that web browsers recognize and visualize. The textual tags are used to assign special properties to the text. Therefore if fragments of the text emerge between two tags (for instance `<b>` and `</b>`), the portion of the included text assumes such tags. Textual tags are the core of the web-page representation method proposed in this paper. Some textual tags, such as those that indicate the page title (`<title>...</title>`) or those that emphasize parts of the text (`<b>...</b>`, `<u>...</u>`, `<em>...</em>`, `<i>...</i>`, and `<strong>...</strong>`), are selected to compute the relevance of each word in a web page. In addition to the criteria derived from the tags above, there are other "classical" attributes that could be considered to compute word relevance, such as *word position* and *word frequency* in the text. Other attributes such as *meta* tags are not considered, as they are not widespread [11]. On the other hand, statistical analyses [12] have proved the difficulty of finding an optimal analytical function that adequately combines all the variables extracted variables.

The fundamental cue is that often a variable evaluates the importance of a word only when it appears combined with another variable. For example, it has been demonstrated that the title does not always describe page content, as in many cases it is the result of an automatic process. The opposite happens with emphasis, since emphasis is an operation consciously performed by the user when designing the web page. Therefore, a word visible in the title will really be relevant when it also appears emphasized or with a high appearance frequency in the text [12].

### 2.1 Definition of Criteria and Variables

The definition of the selected variables follows from the above-mentioned criteria is shown in Table 1.

**Table 1.** Definitions of the variables taken into account in word relevance

Frequency of a Word in the Page	Frequency of a Word in the Title	Word Emphasis Frequency
$c_f(i) = n_f(i) / N_{max}^{page}$	$c_t(i) = n_f(i) / N_{max}^{title}$	$c_e(i) = n_e(i) / N_{max}^{emph}$

Where  $N_{max}^{page}$  is the occurrence of the most frequent word in the page,  $N_{max}^{title}$  the occurrence of the most frequent word in the title, and  $N_{max}^{emph}$  the occurrence of the most emphasized word. Notice that all these variables are normalized and both title and emphasis frequencies take their values in the  $[0,1]$  interval, whereas word frequency in a document never has a 0 value, so its universe of discourse is defined in the  $(0,1]$  interval. The membership-function definitions of the linguistic labels of the variables considered as fuzzy sets can be found in [10]. In addition to these variables, a word-position criterion is always considered. Therefore, in order to compute the relevance of a word from the position variable, the web page is split into four parts to characterize the fact that often users structure the text so that the first and the last lines are more relevant than those in the middle. This fact becomes more significant the longer the text is. The position variable for a word is calculated from the contribution of each word position in the page. The contribution of an occurrence  $o$  of a word  $i$  in a text line  $l$  can be expressed as follows:

$$c_p(i, o) = n_p(i, o) / N_{tot} . \quad (1)$$

Where  $n_p(i, o)$  is the line number of occurrence  $o$  of word  $i$  and  $N_{tot}$  is the total number of text lines in the page. Another fuzzy-rule-based system has been defined to calculate the global position of a word throughout the whole page, and expression (1) is calculated for each occurrence  $o$  of a word  $i$ . Through a fuzzification process, linguistic labels such as INTRODUCTION, BODY and CONCLUSION are assigned at each occurrence  $o$  of the word  $i$  [10]. This procedure captures the gradual behavior exhibited at the transition points of the proposed page partition to be captured. The output of the fuzzy system (the set of IF-THEN rules is described in [10]) is two linguistic labels, STANDARD and PREFERENTIAL, that represent whether or not a word  $i$  belongs, in overall terms, to the favored parts of the text.

## 2.2 The Output of the Fuzzy Knowledge Based System: The Relevance of a Word

The output variable of the fuzzy system formulated to derive the relevance of a word in a hyperlink text has been defined by means of five linguistic labels (see membership functions in Fig 1). The set of all linguistic labels covers the whole range of possible values for a fuzzy variable and has been selected by a human expert scoring each text word by its degree of relevance in the text. The fuzzy IF-THEN rule based system allows for the fusion of the fuzzy variables. These variables have been defined from a knowledge extraction process that accounts for both a previous statistical study [12] and human common sense in the manual performance of the classification tasks. On the other hand, the following premises were considered to design the fuzzy rule based system [10] that derives the word relevance:

- a) The lack of a word in the title indicates either that the page has no title or that the title has no meaning.
- b) Position variable is defined from a criterion to get higher values the longer the page is.

- c) A word with a high frequency means that the word is a “joker,” in the sense that its meaning is not discriminated and thus it can be used in several contexts with different meanings. Notice that the stop-word elimination process does not remove all words whose meaning is unclear.
- d) A word can be non-emphasized simply because no words are emphasized in the whole web page.

Fuzzification of variables, inference and defuzzification processes are performed by means of a tool previously developed at the IAI-CSIC, namely FuzzyShell<sup>1</sup> [13]. The inference engine is defined by means of a center-of-mass (COM) algorithm that weights the output of each rule in the knowledge base with the truth degree of its antecedent. In the approach discussed herein, the output is a linguistic label with an associated value related to the specific word’s relevance in the page.

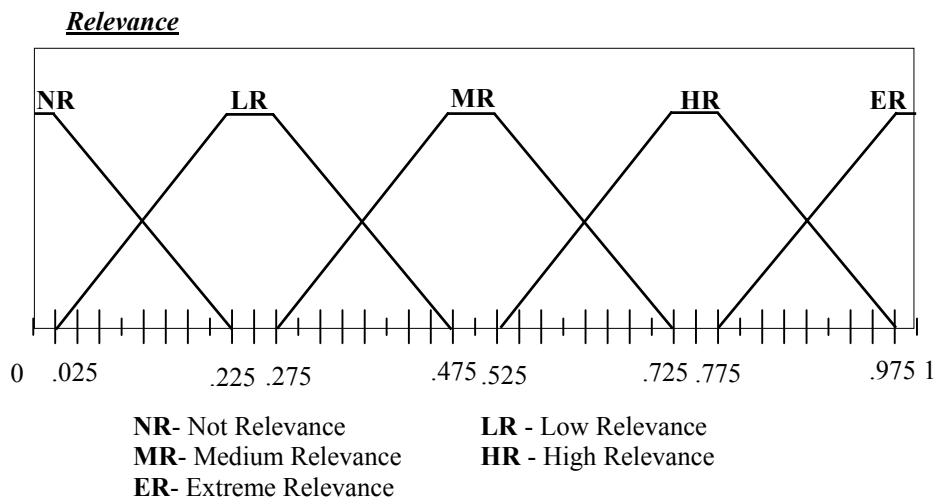


Fig. 1. Membership functions of the output variable *relevance*

### 3 The Supervised Learning Process

In the approach proposed herein, it is assumed that a web-page class is described by a set of fuzzy rules obtained through a supervised learning process. In addition, it shall be accepted herein from this point on that a feature vector, i.e. a two-dimensional vector wherein the first components store the words of the page and second components store the relevance of each word, represents a web page. The first stage in the learning process involves the generation of the corpus, that is, a vector that holds all words contained in all feature vectors of the training set. The relevance for

---

<sup>1</sup> FUZZYSHELL is a CSIC registered trademark 1643983.

each of these words is the attributes under the conditions of the fuzzy classification rules . Consequently, the fuzzy-rule base is composed of a set of rules such that:

**IF** Relev\_word<sub>1</sub> is V<sub>1</sub> **AND** Relev\_word<sub>2</sub> is V<sub>2</sub> **AND**...**AND** Relev\_word<sub>m</sub> is V<sub>m</sub>  
**THEN** Class<sub>1</sub>

Where V<sub>1</sub>, V<sub>2</sub>,... and V<sub>m</sub> are one of the following linguistic values (NR, LR, MR, HR, and ER) defined by trapezoidal membership functions (Fig. 1). The learning process has been demonstrated in the descriptor extraction of two classes, *Medicine-Pharmacology* and *Aerospace-Technology*. The classes were previously analyzed [10] and this enables to contrast the performance of the proposed approach with the former analytical method. The *Medicine-Pharmacology* class is composed of 77 web pages, and the *Aerospace-Technology* class has 101 web pages. The two sets were manually obtained from an expert decision.

Finally, the development of a Genetic Algorithm (GA) for rule discovery involves a number of nontrivial design decisions, such as individual representation, operators, population initialization, selection methods and fitness function. All these decisions are fully described in the paragraphs below.

### 3.1 Individual Representation: Encoding the Fuzzy-Rule Antecedent

In the approach proposed herein, each individual of the GA population represents a set of classification rules, i.e. an entire candidate solution. In the GA context, this rule-inference approach is known as the Pittsburgh approach [14]. The selected codification of an individual involves a fixed-length binary string where bits are grouped in as many five-bit groups as there are words in the corpus. The encoding of an individual is illustrated in Table 2. An individual is encoded as a set of  $m$  conditions where  $m$  is the number of classification attributes, and since each attribute represents the relevance of a word in the corpus,  $m$  is the number of corpus elements. Note that this model implicitly assumes a positional encoding of attributes. In other words, in each individual the first condition refers to the first classification attribute, the second condition refers to the second attribute, and so on. Given this positional convention, the attribute name (in the current case, the word), need not be encoded. In addition, there is an implicit logical AND operator connecting the conditions encoded into the genotype, which is explicitly shown in the decoded rule antecedent. Lastly, each value for each attribute (NR, LR, MR, HR and ER; see Fig. 1) is encoded in a single bit; thus, for each bit of the attribute *relevance\_word<sub>i</sub>* , if the bit is set at "1," the corresponding attribute value would be included in the rule condition, and otherwise it would not. Therefore, keeping in mind the five linguistic labels defined for relevance, the example displayed in Table 2 can be decoded as the antecedent:

(Relev\_word<sub>1</sub> is NR OR HR OR ER) AND (Relev\_word<sub>3</sub> is HR OR ER) AND...AND (Relev\_word<sub>m</sub> is MR). Now then, when two or more linguistic labels are activated in a condition, the antecedent can be split in two or more rules, hence avoiding the definition of the OR operator, as follows:

(Relev\_word<sub>1</sub> is NR) AND (Relev\_word<sub>3</sub> is HR) AND...AND (Relev\_word<sub>m</sub> is MR)  
(Relev\_word<sub>1</sub> is NR) AND (Relev\_word<sub>3</sub> is ER) AND...AND (Relev\_word<sub>m</sub> is MR)

(Relev\_word<sub>1</sub> is HR) AND (Relev\_word<sub>3</sub> is HR) AND...AND (Relev\_word<sub>m</sub> is MR)  
 (Relev\_word<sub>1</sub> is HR) AND (Relev\_word<sub>3</sub> is ER) AND...AND (Relev\_word<sub>m</sub> is MR)  
 (Relev\_word<sub>1</sub> is ER) AND (Relev\_word<sub>3</sub> is HR) AND...AND (Relev\_word<sub>m</sub> is MR)  
 (Relev\_word<sub>1</sub> is ER) AND (Relev\_word<sub>3</sub> is ER) AND...AND (Relev\_word<sub>m</sub> is MR)

**Table 2.** Encoding the antecedent of a rule

Condition 1	Condition 2	Condition 3...	Condition i ...	Condition m
Relev_word <sub>1</sub>	Relev_word <sub>2</sub>	Relev_word <sub>3</sub>	Relev_word <sub>i</sub>	Relev_word <sub>m</sub>
1 0 0 1 1	1 1 1 1 1	0 0 0 1 1	0 0 0 0 0	0 0 1 0 0

Whenever all bits of an attribute are simultaneously set at 1, the corresponding rule condition is not included in the rule antecedent, since the value of the corresponding attribute is irrelevant to determining whether or not a data instance satisfies the rule antecedent. The learning of each class is accomplished in sequence; in other words, in step one the classifier of the *Medicine-Pharmacology* class is achieved, and then the fuzzy-rule set for the *Aerospace-Technology* training examples is built in a similar way. The consequent therefore need not be encoded in the genotype.

### 3.2 Selection Method and Operators

The initial population was randomly generated. Selection is not a problem for the Pittsburgh approach, since in this approach an individual corresponds to a rule set. Hence, a conventional selection method such as a proportional selection (roulette-wheel selection) [14] was used. Interestingly, the positional convention adopted herein simplifies the action of generic operators such as crossovers. By choosing the same crossover points in both parents, the corresponding genetic material can be swapped directly between two individuals without any risk of producing invalid offspring of any sort, such as an individual (rule antecedent) with a duplicate attribute. The two-point crossover was selected and applied with a probability of 0.5. For the mutation operator, a bit-by-bit mutation was used with a probability equal to 0.01.

### 3.3 Fitness Function

In addition to selecting a good representation, it is important to define a good payoff function. Moreover, a classification-rule set's performance in terms of predictive accuracy can be summarized by a contingency matrix [15]. In the current case, this can be accomplished with the 2x2 contingency matrix shown in Fig. 2. The labels in each quadrant of the matrix have the following meaning:

TP = Summarization of the degree of membership in the antecedent for all instances that have class *c*;

FP = Summarization of the degree of membership in the antecedent for all instances that do not have class *c*;

FN = Summarization of the degree of membership in (not (the antecedent)) for all instances that have class *c*;

TN = Summarization of the degree of membership in (not (the antecedent)) for all instances that do not have class  $c$ .

		Actual Class	
		c	not c
Predicted Class	c	TP	FP
	not c	FN	TN

**Fig. 2.** Contingency matrix for a classification-rule set

Note that one could work with a crisp confusion matrix by defuzzifying a rule when the fitness is computed [16]. In the current case, once the degree of matching between the rule antecedent and a data instance is computed, there are two possibilities: (a) If this degree is greater than or equal to 0.5, then the instance satisfies the rule antecedent in a crisp sense; that is, the instance will contribute a value of 1 for either the TP or the FP cell of the confusion matrix, depending on the instance class. (b) If the rule antecedent is not satisfied in a crisp sense, the instances will contribute a value of 1 for either the FN or the TN cell of the confusion matrix, depending on the instance class.

Therefore the fitness function that evaluates the quality of the rule set with respect to predictive accuracy is defined as follows:

$$\text{Fitness} = \text{true\_positive\_rate} \times \text{true\_negative\_rate} = \frac{TP}{(TP + FN)} \times \frac{TN}{(TN + FP)}. \quad (2)$$

## 4 Results

The number of words in the corpus is a function of the number of examples examined (“Ex.” in Table 3) to build the corpus and the length of the feature vector. The size of the corpus thus generated has been analyzed from three points of view (see Table 3): 1) all feature-vector components (Ct), 2) ten components only (C10) and 3) five components. It is important to mention that corpus size decreases considerably with feature-vector size without losing practically any high-relevance words. In fact, for the *Medicine-Pharmacology* class, some words that appear in the corpus generated from 36 examples and the first five components of the feature vector are: “psychiatric,” “pharmacology,” “preclinical,” “clinical,” “drug,” “medication,” “therapy,” “pharmacy,” “medical,” “pharmacotherapy” and “toxic.” In the case of the *Aerospace-Technology* class, some words that appear in the corpus generated from 50 examples and the first five components of the feature vectors are: “space,” “tool,” “aeronautics,” “aerospace,” “aerodynamics,” “flight,” “mechanisms,” “spacecraft,”

“mechanics,” “engineering,” “shuttle,” “fluid,” “dynamics,” “laboratory” and “technology.”

**Table 3.** Corpus size according to the number of feature-vector components selected.

<i>Medicine-Pharmacology</i>				<i>Aerospace-Technology</i>			
Ex.	Ct	C10	C5	Ex.	Ct	C10	C5
77	5546	448	259	101	8255	666	356
38	3285	292	155	50	5456	365	200

Finally, the learning process was performed using 40 examples for each class as input data. Training examples were randomly selected. Next a classification process was carried out for the rest of the examples in order to test the performance of the fuzzy-rule base obtained in the learning step. Some important preliminary results of the comparative study of the proposed fuzzy system versus previous work [10] using a Naive Bayes classifier are shown in Table 4.

**Table 4.** . Main results of the comparative study of the Fuzzy vs. the Naive Bayes classifier

Relevance/ Classifier	Medicine- Pharmacology (%)		Aerospace-Technology (%)		Mean Values (%)	
	Successes	Failures	Successes	Failures	Succes.	Failures
FUZZY / Naive Bayes	84.21	15.79	89.47	10.53	86.84	13.16
FUZZY/ FUZZY	75.05	24.95	72.33	27.67	73.69	26.31

## 5 Conclusion

This paper presents a novel approach to web-page classification. The proposed method offers two major issues. The first is a two-dimensional vector where the first component is words extracted from the textual component of the page and the second component is numeric values that evaluate each word’s relevance in the web page’s text. To calculate word relevance, a fuzzy model is proposed that takes into account some characteristics of HTML language and other more “classic” attributes often used in information retrieval.

The second issue is the use of two web-page sets that exemplify two specific classes to describe a supervised learning method that enables a set of fuzzy rules to be generated to explain the input data sets. Preliminary experiments show that the approach proposed herein is suitable for handling the ambiguity inherent in web-page classification, although wider experimentation is necessary in order to profit from the capacities of a GA search. In the near future improvements should be considered such as a better population-initialization method and a fitness function that accounts for the size of the rule set and penalizes longer rule sets.

## References

1. UNCTAD E-Commerce and development report 2002. Report of the United Nations Conference on Trade and Development. United Nations, New York and Geneva (2002).
2. Gudivada, V.N., Raghavan, V.V., Grosky, W.I., and Kasanagottu, R.: Information retrieval on the World Wide Web. *IEEE Internet Computing*, September-October (1997) 58-68.
3. Chen, H. and Dumais, S.T.: Bringing order to the Web: automatically categorizing search results. *Proceedings of the CHI'00, Human Factor in Computing Systems*, Den Haag, New York, US. ACM Press (2000) 145-152.
4. Salton, G., Wong, A., and Yang, C.S.: A vector space model for information retrieval. *Communications of the ACM*. 18-11 (1975) 613-620.
5. Baeza-Yates, R. and Ribeiro-Neto, B.: *Modern information retrieval*. ACM Press Books, Addison-Wesley (1999).
6. Kosala, R. and Blockeel H.: Web mining research: a survey. *ACM SIGKDD Explorations*. 2-1 (2000) 1-15.
7. Koller, D. and Sahami, M.: Toward Optimal feature selection. *Proceedings of the Thirteenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA (1996) 284-292.
8. Henzinger, M.: Link analysis in web information retrieval. *Bulletin of the Technical Committee on Data Engineering*. 23-3 (2000) 3-8.
9. Yang, Y.: A study of approach to hypertext categorization. *Journal of Intelligent Information Systems*. 18-2/3 (2002) 219-241.
10. Ribeiro, A., Fresno, V., Garcia-Alegre, M.C., and Guinea, D.: A fuzzy system for the web representation. *Intelligent Exploration of the Web. Studies in Fuzziness and Soft Computing*. Szczepaniak, P.S., Segovia, J., Kacprzyk, J., and Zadeh, L.A. Editors. Physica-Verlag, Berlin Heidelberg New York (2003) 19-37.
11. Pierre, J.M.: On the automated classification of web sites. *Linköping Electronic Articles in Computer and Information Science*. Linköping University Electronic Press Linköping, Sweden. 6 (2001).
12. Fresno V. and Ribeiro.: A feature selection and dimensionality reduction in web pages representation. *Proceedings of the International Congress on Computational Intelligence: Methods & Applications*. Bangor, Wales, U.K. (2001) 416-421.
13. Gasós J., Fernández P.D., García-Alegre M.C., García Rosa R.: Environment for the development of fuzzy controllers. *Proceedings of the International Conference on AI: Applications & N.N.* (1990) 121-124.
14. Michalewicz Z.: *Genetic Algorithms + Data Structures = Evolution Programs*. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996).
15. Freitas, A.A.: *Data mining and knowledge discovery with evolutionary algorithms*. Natural Computing Series. Springer-Verlag, Berlin Heidelberg New York (2002).
16. Dasgupta, D. and Gonzales, F.A.: Evolving complex fuzzy classifier rules using a linear tree genetic representation. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'2001)*. Morgan Kaufmann (2001) 299-305.