

# A Prediction System for Cardiovascularity Diseases using Genetic Fuzzy Rule-Based Systems <sup>\*</sup>

O. Cordon <sup>1</sup>, F. Herrera<sup>1</sup>, J. de la Montaña<sup>2</sup>, A.M. Sánchez<sup>3</sup>, and P. Villar<sup>3</sup>

<sup>1</sup>Dept. of Computer Science and A. I., Univ. of Granada, 18071 - Granada, Spain,  
ocordon,herrera@decsai.ugr.es

<sup>2</sup>Lab. of Nutrition and Food Science, University of Vigo, 32004 - Ourense, Spain,  
jmontana@uvigo.es

<sup>3</sup>Dept. of Computer Science, University of Vigo, 32004 - Ourense, Spain,  
amlopez,pvillar@uvigo.es

**Abstract.** In this paper we present a fuzzy rule-based system to predict cardiovascularity diseases. The input variables of the system are the most influent factors for that type of diseases and the output is a risk prediction of suffering from them. Our objective is to get an accurate prediction value and a system description with a high degree of interpretability. We use a set of examples and a design process based on genetic algorithms to obtain the components of the fuzzy rule-based system.

## 1 Introduction

Cardiovascularity diseases are the main cause of mortality in “western countries”. Their prediction is a very complex problem because they are influenced by many factors. The most important of these are diet, age, genetic predisposition, smoking, sedentary life, etc. The development of a cardiovascularity disease takes long time before the first symptoms appear and many times it is too late for the patient. So, it is important to take an adequate preventive action to identify and modify the risk factors associated.

In this contribution, we use a fuzzy rule-based system (FRBS) as a means to determine a prediction to suffer from a cardiovascularity disease. As it is difficult for an expert to design the FRBS due to its complexity, we derive it from a learning process using numerical information. In this paper, we have used a genetic algorithm as learning mechanism, so dealing with a genetic fuzzy rule-based system [6]. Moreover, an important objective of this work is to obtain models that can be interpretable. That is the aim for using FRBS.

This paper is organized as follows. Sections 2 and 3 show some preliminaries about cardiovascularity diseases and FRBSs, respectively. Section 4 presents the problem description. The learning method used to obtain the FRBS is briefly described in Section 5. Section 6 presents some experimental results as well as the complete description of a simple FRBS for cardiovascularity disease prediction. Finally, in Section 7, some conclusions are pointed out.

---

<sup>\*</sup> This research has been supported by CICYT PB98-1319

## 2 Cardiovascularity diseases

The main factors that influence the appearance of a cardiovascularity disease [8, 9] are shown next:

- **Diet.** From the fifties, epidemiological studies have proven the existence of a direct relationship between the amount and type of fatty consumed and the serum levels of *cholesterol* and *triglycerides*, that are considered as the most important factors in the development of cardiovascularity diseases. The cholesterol total amount is the sum of the cholesterol amount associated to the three types of *lipoproteins* more abundant in the blood: very low density lipoproteins (VLDL), low density ones (LDL) and high density ones (HDL). It is useful to distinguish between the cholesterol associated to the LDL (*LDL-cholesterol*) and to the HDL (*HDL-cholesterol*). The LDL-cholesterol is a clear risk factor to suffer from a cardiovascularity disease. On the other hand, the HDL-cholesterol is good to prevent it due its antiatherogenic quality. Diet plays a very important role in the serum levels of any type of cholesterol. The main aspect is the amount of fatty in the diet and the type of fatty acids present in the blood. There are three main types of fatty acids:
  - Saturated Fatty Acids (SFA), that increase the cholesterol levels found in the blood, specially the LDL-cholesterol.
  - Monounsaturated Fatty Acids (MUFA), that are conferred a neutral or slightly beneficial effect over the cholesterol levels.
  - Polyunsaturated Fatty Acids (PUFAS), that origin a reduction of the cholesterol concentration, specially of the LDL-cholesterol.
- **Hypertension.** High values of the diastolic blood pressure predispose to a heart attack and other cardiovascularity diseases.
- **Smoking.** The tobacco consum (specially cigarettes) constitutes an important risk factor for the development of cardiovascularity diseases.
- **Obesity.** The obesity is a negative factor for the health. Some studies connect the weight increase with a progressive increment of the serum levels of cholesterol and triglycerides. The degrees of obesity are classified through the Body Mass Ratio (BMR) that is defined as the division of the weight of an individual (in Kgs.) by its square height (in metres).
- **Sedentary.** A slight physical activity produces positive effects over the cardiovascularity system. On the other hand, a sedentary life increases the obesity and the amount of LDL-cholesterol.
- **Age.** Unlike the previous factors, age can not be modified, but it is very important as it can directly affect to them.

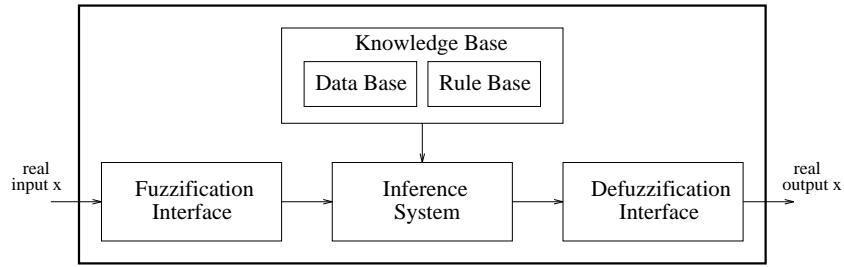
## 3 Fuzzy Rule-Based Systems

FRBSs [13] constitute an extension of classical rule-based systems, as they deal with fuzzy rules instead of classical logic rules. In this approach, fuzzy IF-THEN rules are formulated and a process of fuzzification, inference and defuzzification

leads to the final decision of the system (see Fig. 1). The FRBS is then considered as an approach used to model a system making use of a descriptive language based on Fuzzy Logic with fuzzy predicates. The fuzzy rules used –also called linguistic rules– have the following structure [10]:

$$\text{IF } X_1 \text{ is } A_1 \text{ and } \dots \text{ and } X_n \text{ is } A_n \text{ THEN } Y \text{ is } B_i$$

with  $X_1, \dots, X_n$  and  $Y$  being the input and output linguistic variables, respectively, and  $A_1, \dots, A_n$  and  $B$  being linguistic labels, each one of them having associated a fuzzy set defining its meaning.



**Fig. 1.** Generic structure of a descriptive Fuzzy Rule-Based System

The main characteristic of this type of fuzzy systems is the interpretability of the resulting model. The knowledge base of the FRBS is composed of a collection of rules together with a description of the semantics of the linguistic terms contained in these rules. Of course, if the number of rules is excessively high, the global interpretability of the model decreases, although it is possible to perform a local interpretation for the output of a given input data pair analysing only the rules fired for that input data pair.

Although sometimes the fuzzy rules can be directly derived from expert knowledge, different efforts have been made to obtain an improvement on the system performance by incorporating learning mechanisms guided by numerical information to define them, the Rule Base (RB), and/or the membership functions, the Data Base (DB), associated to them. Most of them have focused on the RB learning, using a predefined DB [3, 12]. This operation mode makes the DB have a significant influence on the FRBS performance [1, 4].

The usual solution to improve the FRBS performance by dealing with the DB components involves a tuning process of the preliminary DB definition once the RB has been derived [1], maintaining the latter unchanged. In opposite to this *a posteriori* DB learning, there are some approaches that learn the different DB components *a priori* [4, 5], thus allowing it to both adjust the membership functions and learn the optimal number of linguistic terms for each variable. This will be the approach followed in this paper to design the FRBS.

## 4 Building our Prediction System

The first step to design a prediction system for cardiovascularity diseases is to decide which variables will be considered. If all the risk factors were taken into account, a very large set of example data pairs would be needed and it would be difficult to get an FRBS with good performance and a high degree of interpretability, i.e. with a low number of rules and a low number of labels per variable.

For that reason, we only consider the most influent risk factors: the different cholesterol levels, the triglycerides level and the age. The remaining factors are individual habits that the experts only consider for a small increment or decrement of the risk prediction associated to the main risk factors. The most important of these “secondary” factors have been aggregated into a single variable called *habits*. Therefore, our system will have six input variables and one output variable that are described as follows:

- *Cholesterol*: Total cholesterol level present in the blood (in *mg/dl*). Range: [100 – 350]
- *LDL-cholesterol*: Serum LDL-cholesterol level (in *mg/dl*). Range: [100 – 210]
- *HDL-cholesterol*: Serum HDL-cholesterol level (in *mg/dl*). Range: [10 – 200]
- *Triglycerides*: Triglycerides level present in the blood (in *mg/dl*). Range: [140 – 160]
- *Age*: Age of the individual (in years). Range: [20, 80]
- *Habits*: This variable takes values in  $\{1, \dots, 48\}$ . The higher the value, the worse the habits (from the point of view of the disease prevention). A value in the range  $\{1 - 12\}$  indicates beneficial factors that reduce the risk of suffering a cardiovascular disease. Values in  $\{13 - 24\}$  represent habits that can be considered “neutral” over the risk prediction. The range  $\{25 - 37\}$  indicates slightly damaging habits. Finally, values in  $\{38 - 48\}$  indicate a dangerous increment of the risk of suffering a cardiovascular disease. The selected habits are shown next:
  - High consum of PUFAS (greater than 33 *gr/day*)
  - High consum of SFA (greater than 33 *gr/day*)
  - Body Mass Ratio (BMR), defined by the relation between the weight of an individual (in kgs.) and the square of its height (in meters). Two ranges are considered:  $20 < BMR < 30$  and  $BMR > 30$
  - Habitual smoker (more than 10 cigarettes per day)
  - Cholesterol consumed per day (*cholest./day*) in milligrams. Three ranges are considered:  $cholest./day < 200mgs$  ,  $200mgs < cholest./day < 300mgs$  and  $cholest./day > 300mgs$
  - Phisycal activity. It implies to make daily exercises, even if they are slight.

Table 1 shows the correspondency between each value of this variable and the concrete habits of the individual.

- *Risk prediction*: The output variable is a numeric real value ( $[0 - 10]$ ) that indicates an estimation of the risk to suffer from a cardiovascularity disease. The higher the value, the higher the risk.

**Table 1.** Variable *habits*: correspondency of the 48 values with the habits considered

Value	high consum PUFAS	high consum SFA	BMR between 20-30	BMR greater 30	Smoker	< 200 mgs. choles./day	200 – 300 mgs. choles./day	> 300 mgs. choles./day	Phisi- cally activ
1	x		x			x			x
2	x		x		x	x			x
3	x		x			x			
4	x		x		x	x			
5	x		x				x		x
6	x		x		x		x		x
7	x		x				x		
8	x		x		x		x		
9	x		x					x	x
10	x		x		x			x	x
11	x		x					x	
12	x		x		x			x	
13	x			x		x			x
14	x			x	x	x			x
15	x			x		x			
16	x			x			x		
17	x			x			x		x
18	x			x	x		x		x
19	x			x	x		x		
20	x			x	x	x			
21	x			x				x	x
22	x			x	x			x	x
23	x			x				x	
24	x			x	x			x	
25		x	x			x			x
26		x	x		x	x			x
27		x	x			x			
28		x	x		x	x			
29		x	x				x		x
30		x	x		x		x		x
31		x	x				x		
32		x	x		x		x		
33		x	x					x	x
34		x	x		x			x	x
35		x	x					x	
36		x	x		x			x	
37		x		x		x			x
38		x		x	x	x			x
39		x		x		x			
40		x		x	x	x			
41		x		x			x		x
42		x		x	x		x		x
43		x		x			x		
44		x		x	x		x		
45		x		x				x	x
46		x		x	x			x	x
47		x		x				x	
48		x		x	x			x	

Our method uses numerical information for the learning process. Unfortunately, it is very difficult to obtain data from real patients. It would imply frequent studies of the biochemical parameters of many healthy people (from the cardiovascular diseases point of view) with different serum levels, age and habits, and to wait (sometimes many years) for a possible development of a cardiovascular disease. So, the examples have been generated *ad hoc* by an expert trying to produce an uniformly distributed set covering a great range of possible situations. The obtained set is composed of 2594 data pairs and it has been randomly divided into two subsets, a training data set with 2335 elements (90%) and a test data one with 259 elements (10%).

## 5 Learning process to derive the FRBS

We use the genetic fuzzy rule-based system proposed in [5] to learn the KB of the FRBS associated to our prediction system for the risk to suffer a cardiovascularity disease. We will refer that process as **GFS** (Genetic Fuzzy System). **GFS** is composed of two methods with different goals:

- A Genetic Algorithm (GA) [11] to learn the DB that allows us to define:
  - The number of labels for each linguistic variable (granularity).
  - The variable domain (working range), allowing a brief enlargement of the initial domain.
  - The form of each fuzzy membership function (triangular-shaped) in non-uniform fuzzy partitions using a non-linear scaling function.
- A quick *ad hoc data-driven method* [3] that derives the RB considering the DB previously obtained. It is run from each DB definition generated by the GA. In this paper we use the inductive method proposed in [12].

There are three steps that must be done to evaluate each chromosome:

1. Generate the fuzzy partitions (DB) for all the linguistic variables using the information contained in the chromosome.
2. Generate the RB by running the fuzzy rule learning method considering the DB obtained.
3. Calculate the fitness function. First the Mean Square Error (MSE) over the training set is calculated from the KB obtained (genetically derived DB + RB). This value will be used as a base of the chromosome fitness value:

$$MSE = \frac{1}{2|E|} \sum_{e_l \in E} (ey^l - S(ex^l))^2$$

with  $E$  being the example set,  $S(ex^l)$  being the output value obtained from the FRBS when the input variable values are  $ex^l = (ex_1^l, \dots, ex_n^l)$ , and  $ey^l$  being the known desired output value.

In order to improve the generalization capability and the interpretability of the final FRBS, we will lightly penalize FRBSs with a high number of rules to obtain more compact linguistic models. Therefore, once the RB has been

generated and its MSE over the training set ( $MSE_{tra}$ ) has been calculated, the fitness function is calculated in the following way [7]:

$$F_C = \omega_1 \cdot MSE_{tra} + \omega_2 \cdot N\_Rules$$

with  $N\_Rules$  being the number of rules and  $\omega_1, \omega_2$  two weighting factors.

## 6 Experimental results

We have run the **GFS** process with different initial seeds, using various ranges for the granularity values across the interval  $\{2, 9\}$ . **GFS** allows us to obtain FRBSs with a different trade-off between accuracy and interpretability reducing the maximum value for the granularity and changing the values of the fitness function weights (parameters  $\omega_1$  and  $\omega_2$ ). The genetic parameters considered are the following: number of generations=1000, population size=100, crossover probability=0.6, mutation probability=0.1. We have also considered other types of learning methods in order to compare with the results obtained by **GFS**. We run the following methods:

- Linear Regression.
- Neural Networks (NN): A three layer perceptron, using conjugate gradient plus weight decay as learning rule. Different values for the number of units in the hidden levels were considered. The table of results shows results of two NN, the one with the best  $MSE_{tra}$  and the one with the best  $MSE_{tst}$
- A representative process of the usual way to derive an FRBS: The Wang and Mendel’s rule generation method plus a DB tuning process (WM + Tuning). As usual, all the variables have the same granularity. We run the WM method for all the possible numbers of labels considered ( $\{2, \dots, 9\}$ ) and the best results considering the  $MSE_{tra}$  was obtained with nine labels while the best result over the  $MSE_{tst}$  was obtained with four labels. We have used the genetic tuning process proposed in [2] to refine the preliminary DB of both FRBSs once the RB has been derived.

Linear regression obtains models that can not be considered totally interpretables, while the Neural Networks are not interpretables. Both are shown in order to compare the accuracy of the method proposed for modelling the prediction system (**GFS**). The best results obtained are presented in Table 2, which contains the following columns:

- **Method**: Process used to model the prediction system
- **Granularity**: Number of labels per variable (for FRBS learning methods)
- **N\_Rules**: Number of rules of the FRBS RB (for FRBS learning methods)
- **MSE<sub>tra</sub>**: MSE over the training data set
- **MSE<sub>tst</sub>**: MSE over the test data set

As can be observed, many learning methods obtain good results as regards the prediction ability of the resulting model. The best result in  $MSE_{tra}$  has been obtained using a multilayer perceptron with 20 units in the hidden level, although

**Table 2.** Best results obtained

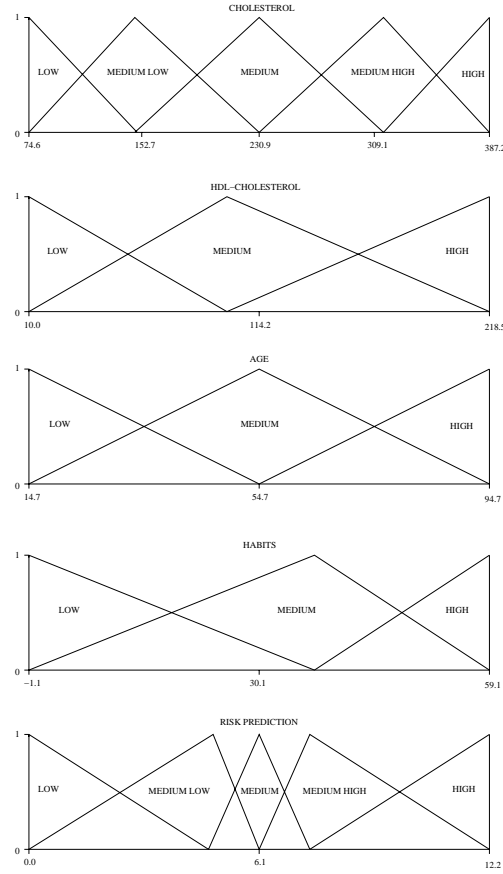
Method	Granularity	N_Rules	$MSE_{tra}$	$MSE_{tst}$
Linear Regression	- - - - -	-	0.0741	0.0683
NN 6-20-1	- - - - -	-	0.0445	0.0691
NN 6-5-1	- - - - -	-	0.0639	0.0645
WM + Tuning	9 9 9 9 9 9	2222	0.0785	2.8523
	4 4 4 4 4 4	913	0.1347	0.1589
GFS	9 5 6 4 5 3 9	319	0.0451	0.0474
	9 8 7 5 4 4 9	133	0.0611	0.0606
	7 2 3 2 3 4 6	43	0.0883	0.0892
	5 4 3 3 2 2 5	31	0.1042	0.0996
	5 3 3 3 3 3 5	23	0.1274	0.1171

there is only a small difference respect to the best result obtained with **GFS**. The best result in  $MSE_{tst}$  has been obtained using the **GFS** process considering the interval  $\{3, \dots, 9\}$  as possible granularity values. So, the prediction ability of the models obtained by **GFS** are enoughly demonstrated. Regarding to the usual process to derive a FRBS (WM + Tuning), the choice of a high number of labels produces good results in  $MSE_{tra}$  but clearly leads to an overfitting as can be observed in the great value obtained for  $MSE_{tst}$ .

The table collects different FRBSs obtained from the **GFS** process, some of them with good results in the  $MSE$  columns and others with low values in the granularity and number of rules. Of course, the latter present greater degrees of interpretability than the former. As said, it is possible to obtain FRBSs with good accuracy or great interpretability by changing the range of the granularity levels and the weighing factors in the fitness function of the GA. The most accurate FRBSs present more rules and a higher granularity level than the most interpretable FRBSs displayed.

In order to show an example of the composition of an FRBS for the problem, the most simple FRBS of Table 2 (FRBS with 21 rules) are described. A typical consequence when a learning method is forced to obtain FRBS with a few rules is the implicit elimination of the input variables with lesser relevance, that is, if one variable has the same label in all the rules, it has not influence in the prediction ability. So, we will ignore two variables in the description of this FRBS (*LDL-Cholesterol* and *Tryglicerides*). Figure 2 shows the DB (fuzzy partitions for all the relevant variables in this specific FRBS including the new domain limits learned by **GFS**). In order to improve the readability of the RB, if two rules only differ in one input label (the remaining input variables and the output variable have the same linguistic term), they are depicted as a single rule including the two different labels connected with the operator OR. Therefore, the RB is composed of the following rules:

- R<sub>1.2</sub>: IF Cholesterol is *HIGH* and HDL-cholesterol is *HIGH* and Age is *MEDIUM* and Habits is (*LOW* or *MEDIUM*) THEN Risk is *MEDIUM*
- R<sub>3.4</sub>: IF Cholesterol is *HIGH* and HDL-cholesterol is *HIGH* and Age is *HIGH* and Habits is (*LOW* or *MEDIUM*) THEN Risk is *MEDIUM*



**Fig. 2.** Fuzzy partitions

- R<sub>5\_6</sub>: IF Cholesterol is *HIGH* and HDL-cholesterol is *MEDIUM* and Age is *MEDIUM* and Habits is (*LOW* or *MEDIUM*) THEN Risk is *MEDIUM HIGH*
- R<sub>7\_8</sub>: IF Cholesterol is *HIGH* and HDL-cholesterol is *MEDIUM* and Age is *HIGH* and Habits is (*LOW* or *MEDIUM*) THEN Risk is *MEDIUM HIGH*
- R<sub>9</sub>: IF Cholesterol is *HIGH* and HDL-cholesterol is *HIGH* and Age is *LOW* and Habits is *MEDIUM* THEN Risk is *MEDIUM*
- R<sub>10\_11</sub>: IF Cholesterol is *MEDIUM HIGH* and HDL-cholesterol is *HIGH* and Age is *MEDIUM* and Habits is (*LOW* or *MEDIUM*) THEN Risk is *MEDIUM LOW*
- R<sub>12\_13</sub>: IF Cholesterol is *MEDIUM HIGH* and HDL-cholesterol is *MEDIUM* and Age is *HIGH* and Habits is (*LOW* or *MEDIUM*) THEN Risk is *MEDIUM*
- R<sub>14\_15</sub>: IF Cholesterol is *MEDIUM HIGH* and HDL-cholesterol is *MEDIUM* and Age is *MEDIUM* and Habits is (*LOW* or *MEDIUM*) THEN Risk is *MEDIUM*
- R<sub>16</sub>: IF Cholesterol is *MEDIUM HIGH* and HDL-cholesterol is *HIGH* and Age is *HIGH* and Habits is *MEDIUM* THEN Risk is *MEDIUM*
- R<sub>17</sub>: IF Cholesterol is *MEDIUM HIGH* and HDL-cholesterol is *HIGH* and Age is *HIGH* and Habits is *LOW* THEN Risk is *MEDIUM LOW*
- R<sub>18\_19</sub>: IF Cholesterol is *MEDIUM* and HDL-cholesterol is *MEDIUM* and Age is *MEDIUM* and Habits is (*LOW* or *MEDIUM*) THEN Risk is *MEDIUM LOW*

- R<sub>20-21</sub>: IF Cholesterol is *MEDIUM* and HDL-cholesterol is *MEDIUM* and Age is *HIGH* and Habits is (*LOW* or *MEDIUM*) THEN Risk is *MEDIUM LOW*
- R<sub>22-23</sub>: IF Cholesterol is *MEDIUM LOW* and HDL-cholesterol is *MEDIUM* and Age is (*MEDIUM* or *HIGH*) and Habits is *LOW* THEN Risk is *MEDIUM LOW*

## 7 Concluding remarks

We have proposed an FRBS to predict the risk of suffering from a cardiovascular disease. The learning process uses a GA for deriving the DB and a simple RB generation method to learn the rules. The FRBS learning process allows us to choose the main characteristic desired for the prediction model: good accuracy or good interpretability. As future works, we will try to design an FRBS to predict the risk taking all the human habits as a base. This new approach could be more interesting for the nutrition specialist in order to advise the patients.

## References

1. Bonissone, P.P., Khedkar, P.S., Chen, Y.T.: Genetic algorithms for automated tuning of fuzzy controllers, a transportation application, Proc. Fifth IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'96) (New Orleans, 1996) 674-680.
2. Cordon, O., Herrera, F.: A three-stage evolutionary process for learning descriptive and approximative fuzzy logic controller knowledge bases from examples, International Journal of Approximate Reasoning 17(4) (1997) 369-407.
3. Casillas, J., Cordon, O., Herrera, F.: COR: A methodology to improve ad hoc data-driven linguistic rule learning methods by inducing cooperation among rules, IEEE Tr. on Systems, Man, and Cybernetics-Part B: Cybernetics (2002). To appear.
4. Cordon, O., Herrera, F., Villar, P.: Analysis and guidelines to obtain a good uniform fuzzy partition granularity for fuzzy rule-based systems using simulated annealing, International Journal of Approximate Reasoning 25(3) (2000) 187-216.
5. Cordon, O., Herrera, F., Magdalena, L., Villar, P.: A genetic learning process for the scaling factors, granularity and contexts of the fuzzy rule-based system data base, Information Science 136 (2001) 85-107.
6. Cordon, O., Herrera, F., Hoffmann, F., Magdalena, L.: Genetic fuzzy systems. Evolutionary Tuning and Learning of Fuzzy Knowledge Bases, (World Scientific, 2001).
7. Ishibuchi, H., Nozaki, K., Yamamoto, N., Tanaka, H.: Selecting fuzzy if-then rules for classification problems using genetic algorithms, IEEE Tr. on Fuzzy Systems 3(3) (1995) 260-270.
8. Mahan, L.K., Scott-Stump, S.: KRAUSE'S Food, Nutrition and Diet Therapy, (W.B. Saunders, 1996).
9. Mann, J.: Diseases of the heart and circulation: the role of dietary factors in aetiology and management, in: J.S. Garrow and W.P.I. James, Eds., Human nutrition and dietetics, (1993) 619-650.
10. Mamdani, E.H.: Applications of fuzzy algorithm for control a simple dynamic plant, Proceedings of the IEEE 121(12) (1974) 1585-1588.
11. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs, (Springer-Verlag, 1996).
12. Wang, L.X., Mendel, J.M.: Generating fuzzy rules by learning from examples, IEEE Tr. on Systems, Man, and Cybernetics 22 (1992) 1414-1427.
13. Zadeh, L.A.: Outline of a new approach to the analysis of complex systems and decision processes, IEEE Tr. on Systems, Man, and Cybernetics 3(1) (1973) 28-44.