Journal of Babylon University/Pure and Applied Sciences/ No.(4)/ Vol.(21): 2013

Breast Cancer Diagnosis Using Genetic Fuzzy Rule Based System

Hussein A.Lafta , Noor Kadhum Ayoob
Science College For Women

Abstract

Breast cancer diagnosis (WBCD) is an important, real-world medical problem. There are different artificial Intelligence techniques try to classify WBCD to help to minimize the errors that might occur when the doctors do not have adequate experience or because of stress . In this work , fuzzy genetic tool is used to present diagnostic system that classify WBCD cases automatically .The system provides two prime features: first, it attain high classification performance ; second, the resulting system consists of a few simple rules, and are therefore interpretable.

Keywords: WBCD ;Fuzzy systems; Genetic algorithms; Breast cancer diagnosis; GFRBS

الخلاصة

يعد تشخيص سرطان الثدي (WBCD) من المشاكل الواقعية المهمة في المجال الطبي . هناك العديد من تقنيات الذكاء الاصطناعي التي حاولت تصنيف WBCD بهدف المساعدة في تقليص الأخطاء التي يمكن أن تقع عندما لا يمتلك الطبيب الخبرة الكافية أو بسبب الإجهاد . في هذا العمل تم استخدام تقنية هجينة تجمع بين المنطق الضبابي و الخوارزميات الجينية لتقديم نظام تشخيص يصنف حالات WBCD أوتوماتيكياً (بالاعتماد على الحاسوب بصورة كلية) . هذا النظام يوفر خاصيتين مميزتين :

الخاصية الأولى : أن النظام يحقق أداء عالي في التصنيف .

الخاصية الثانية : يعتمد النظام على قواعد ضبابية بسيطة و قليلة و لهذا فهو يقدم تفسيرا لآلية صنع القرار في النظام.

1. Introduction

Breast cancer is the most dangerous disease that threaten women and even men all over the world. After lung cancer breast cancer is the second leading cause of cancer death in women. Over the past few decades, Researchers have been tried to present computerized diagnostic tools to help the physician in diagnosing this cancer .

A good computer-based diagnostic system should possess two important features:
1-The system should attain the highest possible performance providing a numeric value that represents the degree to which the system is confident about its response.
2- Interpretability i.e. the system gives explanation about how the decision is made.
In this work, fuzzy logic and genetic algorithm are used to produce automatically breast cancer diagnosis systems. Fuzzy logic makes the system interpretable while the genetic algorithm makes production of fuzzy systems automatic. In the next two sections, a brief overview of fuzzy systems and genetic algorithms is presented. In Section 4 describes the WBCD problem, which is the main focus of this work. This is followed by an explanation of GFRBS (fusion between GA and fuzzy logic). Section 6 describes in details GFRBS that is used to solve WBCD problem in this work. In Section 7, the results obtained by the system are displayed, followed by conclusion in Section 8.

2. Fuzzy systems

This section explain Fuzzy logic concepts briefly . A more in depth explanation can be found in [Lee, 1990] [Zadeh, 1965].

2.1. Linguistic variables

A linguistic variable [Jain and Abraham, 2004] is defined by its name and its value which is called fuzzy values or labels ,each fuzzy label has a membership function that assign membership degree $\mu_{Label}(x)$ to a crisp element x that is belong to a predefined range of discrete or continuous values , this range known as universe of discourse (UOD) or simply universe.

In classical set theory an element must either belong or not belong to the set and there is no possibility to partial belonging. In contrast, in fuzzy set theory, elements can belong by a certain degree (membership degree) The value of membership degree ranges from 0 to 1 .Let x be an element belong to UOD called X , and A is a fuzzy label :

- When the element x does not belong to the fuzzy label A , the degree of membership ($\mu_A(x)$) is 0.
- When the element x certainly belong to the fuzzy label A , the degree of membership ($\mu_A(x)$) is 1.
- When the element x partially belong to the fuzzy label A , the degree of membership ($\mu_A(x)$) is in the interval [0,1].

2.2 Fuzzy logic Operations

Let A and B be fuzzy sets with the corresponding membership functions $\mu_A(u)$ and $\mu_B(u)$ in the universe U. The following definitions are given in [Lee , 1990].

- **Union**

The union between two fuzzy sets(A and B) is described by the membership function $\mu_{A\cup B}(u)$ and is defined for all $u \in U$ by:

$$\mu_{A\cup B}(u) = \max(\mu_A(u), \mu_B(u))$$

- **Intersection**

The intersection between two fuzzy sets(A and B) is described by the membership function $\mu_{A\cap B}(u)$ and is defined for all $u \in U$ by:

$$\mu_{A\cap B}(u) = \min(\mu_A(u), \mu_B(u))$$

2.3. The structure of fuzzy systems

Fuzzy systems use the linguistic variables to make decisions based on fuzzy rules and this is the reason why these systems get a better response compared to systems using crisp values. A basic structure of a fuzzy system can be seen in figure 2. The main components in the fuzzy system are Fuzzification interface, knowledge base, decision making logic and defuzzification interface .

- **Fuzzification**

In this phase, the system read the input data, scale it to fit the appropriate universe and fuzzify the input data to appropriate linguistic variables that can be handled as fuzzy sets. Scaling the input data to map against the universe appropriate for the system can be done by assigning each membership functions with an explicit function [Lee ,1990]. The out put of this phase is called fuzzy input .

- **Knowledge Base**

The knowledge base contains both of membership functions, known as the database , and a set of fuzzy rules, known as the rule base. These fuzzy rules define the connection between input and output fuzzy variables. A fuzzy rule has the form:
if antecedent then consequent
The antecedent is a fuzzy expressions connected by fuzzy operators (and ,or) while consequent is an expression that assign fuzzy value to the output variables

- **Interface Engine**

In this part , the output from the fuzzy rules is combined depending on how the system is to behave .The typical choices for the reasoning mechanism are Mamdani-type, Takagi-Sugeno-Kang (TKS)-type, and singleton-type [Yager, 1994]. The decision-making process is performed by the inference engine using the rules contained in the rule base.

- **Defuzzification**

The purpose of this phase is to translate the current fuzzy output into a crisp value . This can be done by using methods [Van, 1999] such as Center of Gravity or Center of Area (COA) and the mean of maxima (MOM) methods being the most popular [Mendel, 1995], [Yager, 1994].

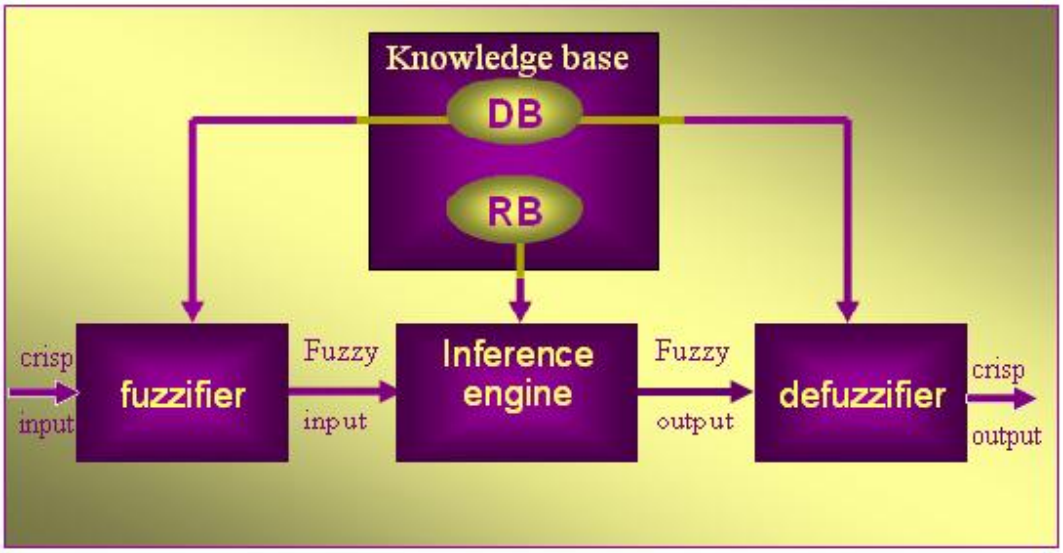

Figure 1 : Basic structure of a fuzzy interface system.

3. Genetic algorithms

A genetic algorithm is an iterative procedure that consists of a population of individuals, each one represented by a finite string of symbols, known as the chromosome which represent a possible solution in a given problem space. Each position in the chromosome is called gene. The value of gene could be binary, real or integer. Chromosomes could be a vector , which is the most common representation , but tree representations, and other representations [Michalewicz, 1996] appear in the recent years. The initial population is generated randomly , A new population is then created from the previous one through several steps [Goldberg, 1989] :

3.1 Evaluation

The fitness of an individual is basically a measure of the individual's ability to solve the problem. How this is done depends on what the problem that is to be solved looks like.

3.2 Crossover / Recombination

Based on the fitness of the individuals they are chosen to be part of a mating pool. crossover is performed with probability pc (the probability of crossover) between two selected individuals, called parents, by exchanging parts of their chromosomes to form two new individuals, called offspring. there are several methods to perform crossover such as 1X , 2X,Ux and so on ,,,

3.3. Mutation

To maintain genetic diversity in the population the newly created individuals go through mutation process in which a simple change in each of the individuals with some (usually small) probability pm to prevent the population from becoming all too similar.

3.4 Next Generation

The fitness value of the newly created individuals are then computed. the new individuals replace the old ones or compare to the old population's fitness values and the individuals with the highest fitness values of both the old and new populations are then chosen to create the next generation. The next generation is generally of equal size as that of the old one. The entire process is then repeated until an acceptable fitness has been reached or until a predefined number of generations have been exceeded.

4. The Wisconsin breast cancer diagnosis problem (WBCD)

In this section the medical diagnosis problem which is the object of this study is presented. Breast cancer is the most common cancer among women. The importance of this problem comes from the fact that it is the major cause of death among women and how the normal cells turn to be cancerous is still unclear. Breast cancer affects men too but rarely .The only way to survive is by early detection. If the cancerous cells are detected before spreading to other organs ,the survival rate for patients is more than 97% (American Cancer society Homepage 2008)

Depending on FNA (Fine Needle Aspiration) test [Mangasarian, 1990] , the University of Wisconsin Hospital for accurately diagnosing breast masses presents data base called The Wisconsin breast cancer diagnosis (WBCD) [Merz, 1996] .The diagnostics in the WBCD database were constructed by specialists in the field. The database contain 699 cases, 16 cases are reported incomplete so they were deleted [setiono, 2000] , the resulting data base has 683 cases in which 239 cases are malignant and 444 cases are benign . Nine attributes are detailed in the figure 2. Each attribute is assigned an integer value between 1 and 10 .

5. Genetic Fuzzy Rule Based Systems (GFRBS)

There are two main approaches to evolve rule systems in the evolutionary algorithm : the Michigan approach and the Pittsburgh approach [Michalewicz, 1996]. A new method has been proposed specifically for fuzzy modeling : the Iterative rule learning approach [Herrera, 1995]. These three approaches are presented below.

5.1. The Michigan approach

This method was first developed by Holland and Retain 1983. Each individual represents a single rule thus the entire population is represented rule base . The rules are competition for the best action to be proposed, and cooperate to form an efficient fuzzy

system. It is difficult to make the decision of which rules are ultimately responsible for good system behavior because of the cooperative and competitive nature .

Attribute No.	Attribute name	Attribute values
1	clump thickness	1 – 10
2	uniformity of cell size	1 – 10
3	uniformity of cell shape	1 – 10
4	marginal adhesion	1 – 10
5	single epithelial cell size	1 – 10
6	bare nuclei	1 – 10
7	bland chromatin	1 – 10
8	normal nucleoli	1 – 10
9	mitoses	1 – 10

No. of cases is 683 , 239 malignant and 444 benign

Figure 2 : WBCD description

5.2. The Pittsburgh approach

In this approach , entire rule base is encoded in the individual. Selection and genetic operators are used to produce new generations of fuzzy systems. This approach allows to include additional optimization criteria in the fitness function, thus affording the implementation of multi-objective optimization. The main drawback of this approach is its computational cost, since a population of fuzzy systems has to be evaluated each generation. This method was first introduced by Smith in 1980.

5.3. The iterative rule learning approach

his approach combines the speed of the Michigan approach with the simplicity of fitness evaluation of the Pittsburgh approach. Like Michigan approach, each individual encodes a single rule. An evolutionary algorithm is used to find a single rule that is the best rule in the population, thus building rule base step by step until an appropriate rule base is built. To prevent the process from finding redundant rules (i.e. rules with similar antecedents), a penalization scheme is applied each time a new rule is added. However , this method can lead to a non-optimal partitioning of the antecedent space.

6. GFRBS for the WBCD problem

In this section describes the parameters of GA and fuzzy system that are used in this work. Subsection 6.1 is devoted to describe Fuzzy system and its parameters that are used in the system. Subsection 6.2 describes GA and its parameters for WBCD problem in this work.

6.1 Fuzzy system parameters

1- **Reasoning mechanism** : singleton-type fuzzy system

2- **Membership functions** :

- The type and the number of Input membership function : two trapezoidal denoted Low and High and one triangular denoted as medium are used (Fig. 3).

- The type and the number of output membership function : two singletons are used to indicate benignity and malignancy. Depending on WBCD , if the case is benign , the value of output singleton is 2 . The value of output singleton is 4 when the case is malignant .

3- **Rules** :

- No. of rules : in this approach the no. of rules are specified by the user but the rules themselves are to be found by the genetic algorithm.
- Antecedents of rules: they are to be found by the genetic algorithm by determining the attributes that participate in the rules and their fuzzy values.
- Consequent of rules: the algorithm finds rules for the benign diagnostic; the malignant diagnostic is an else condition, this mean that there is no need to evolve consequent part in the chromosome. For example :

R1 : if (v3 is Low) and (v7 is Low) and (v8 is Low) then diagnosis is benign
R2 : if (v1 is Low) and (v2 is Low) and (v3 is High) then diagnosis is benign
else diagnosis is malignant

4- **Defuzzification method**: weighted average.

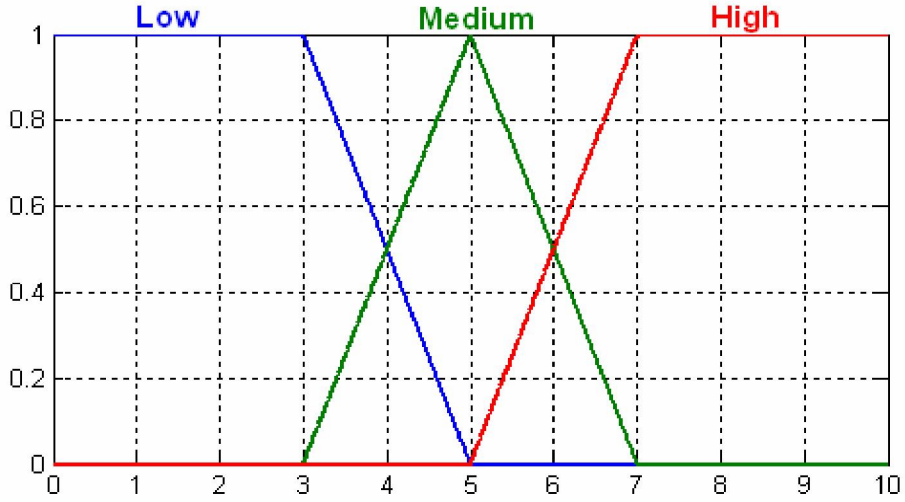

Figure 3 : Membership functions for WBCD attribute

6.2. Genetic algorithm parameters

- **GFRBS type** : Pittsburgh-style structure learning.
- **Encoding method (gene value)** : Hybrid encoding (integer for database and binary for rule base).
- **No. of individual per population (best population size)** = 50 .
- **Crossover type** : UX.
- **Mutation type** : 2m (for database part) and 1m (for rule base part).
- **Selection method** : elitism.
- **Stopping condition** : No. of generations = 50 .

- **Parameters encoded in the chromosome** : for each variable, the beginning of triangular membership (p) , and its base length (d) are encoded in the chromosome. In addition ,for each rule there are 18 genes determine fuzzy value (each value is represented by two genes) for each variable (1 for Low ,2 For High ,3 for Medium , otherwise the variable is not part of the rule) , figure 4 shows an example of chromosome structure. figure (5) show the interpretation of the chromosome shown in figure (4):
- **Chromosome length** = 18 + 18 * (No. of rules).

p_1	d_1	p_2	d_2	p_3	d_3	p_4	d_4	p_5	d_5	p_6	d_6	p_7	d_7	p_8	d_8	p_9	d_9
٣	٤	٤	٢	٣	٤	٦	١	٨	٣	٢	٥	٧	٧	٥	٥	٥	٤

v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
1	1	0	0	3	١	٠	١	0

Figure (4) : Chromosome structure

Database

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
p	3	4	3	6	8	2	7	5	5
d	4	2	4	1	3	5	7	5	4

Rule base

if (v_1 is Low) and (v_2 is Low) and (v_5 is medium) (v_6 is Low) and (v_8 is Low) then
diagnosis is benign
else diagnosis is malignant

Figure (5) : Chromosome interpretation

7. Results

This section describes the results obtained when applying the proposed system described in Section 6. The evolutionary experiments performed fall into three categories, in accordance with the data repartitioning into two distinct sets: training set and test set. The three experimental categories are:

(1) training set contains all 683 cases of the WBCD database, while the test set is empty.

(2) training set contains 75% of the WBCD cases, and the test set contains the remaining 25% of the cases.
(3) training set contains 50% of the WBCD cases and the test set contains the remaining 50% of the cases.

In the last two categories, the choice of training-set cases is done randomly at each run. A total of 150 evolutionary runs were performed, all of which found systems , the results shown in table 1,2, and 3.The best system has four rules and its performance is 97.9502 with an average of variable equal to 4 :

R1 : if (v2 is Low) and (v4 is High) and (v6 is High) and (v9 is Low) then diagnosis is benign

R2 : if (v2 is Low) and (v3 is High) and (v4 is High) and (v5 is High) and (v6 is Low) then diagnosis is benign

R3 : if (v2 is Low) and (v3 is Low) and (v6 is Low) then diagnosis is benign

R4 : if (v1 is Low) and (v3 is Low) and (v5 is Low) and (v8 is High) then diagnosis is benign

Default else diagnosis is malignant

8. Conclusion

In this work WBCD database is classified by using GFRBS approach. The system attains high classification performance and the resulting system has a few simple rules, and are therefore interpretable. Experience shows that the fuzzy-genetic tool is promising approach where such medical diagnosis problems are concerned. As future works, another techniques can be used (for example, fuzzy Petri net or fuzzy networks instead of FRBS) to improve the current system and applying the GFRBS approach to more complex diagnosis problems .

Table 1 : The results of the system when training set is 100% (all cases) and the test set is 0 % (contain no data)

One rule		Two rules		Three rules		Four rules		Five rules	
%	Av	%	Av	%	Av	%	Av	%	Av
97.36	4	97.80	3.5	97.36	3.7	97.95	٤	97.66	4
97.36	4	97.51	4.5	97.36	3.7	97.36	3.2	97.66	4.2
97.36	4	97.36	3	97.36	4	97.22	4.3	97.36	3.4
97.36	4	97.36	3	97.36	4	97.07	3	97.07	3.2
97.36	6	97.36	4	97.36	4	97.07	3.5	97.07	3.4
97.22	5	96.93	3	97.36	4.3	96.93	2.8	96.93	3.4
97.07	5	96.93	3.5	97.36	4.3	96.93	4	96.49	3.2
96.93	4	96.78	2.5	97.36	4.3	96.78	2.3	96.34	2.8
96.49	4	96.78	3	97.36	4.3	96.78	3.8	96.34	3.2
96.19	3	96.49	4	97.07	3.7	96.49	4	95.61	2.6
Max = 97.36 Min =96.19		Max =97.80 Min =96.49		Max =97.36 Min =97.07		Max =97.95 Min =96.49		Max =97.66 Min =95.61	

Table 2 : The results of the system when training set is 75% and the test set is 25 % .

One rule				Two rules				Three rules				Four rules				Five rules			
train	test	%	Av	train	test	%	Av	train	test	%	Av	train	test	%	Av	train	test	%	Av
97.46	97.08	97.36	4	97.07	98.25	97.36	3	97.66	97.08	97.51	3.7	98.05	96.49	97.66	3.3	97.85	95.91	97.37	4
97.27	97.08	97.22	4	97.66	95.91	97.22	3	97.27	97.66	97.36	3	97.66	96.49	97.37	3.8	97.66	95.91	97.22	3.4
97.46	95.91	97.07	4	97.85	94.74	97.07	4	97.27	96.49	97.07	4	97.46	96.49	97.22	2.8	97.66	95.32	97.07	2.6
96.88	97.08	96.93	4	95.90	94.74	95.61	2	96.68	96.49	96.63	4	97.46	96.49	97.22	3	97.46	95.32	96.93	3.2
97.46	94.74	96.78	5	97.27	95.32	96.78	3	97.66	92.98	96.49	3.3	97.46	96.49	97.22	3.5	96.88	95.91	96.93	3
97.46	95.91	97.07	4	97.07	95.91	96.78	3	96.88	95.32	96.49	2.7	96.88	96.49	96.78	3.8	96.48	97.08	96.63	4
96.88	95.91	96.63	3	96.48	97.08	96.63	3	96.88	95.32	96.49	3	96.48	97.08	96.63	3.3	96.88	95.32	96.49	3.2
96.48	95.91	96.34	4	97.07	94.15	96.34	3.5	96.88	94.15	96.19	3.7	97.27	94.15	96.49	3.5	97.27	93.57	96.34	3
96.29	95.32	96.05	4	96.68	97.66	96.93	3.5	96.68	94.74	96.19	2.7	96.88	95.32	96.49	3.8	96.48	95.32	96.19	3.6
95.12	92.98	94.58	3	95.90	92.40	95.02	2	97.27	92.40	96.04	4	97.27	92.40	96.05	3.8	96.29	92.98	95.46	2.8
Max = 97.36 Min = 94.58				Max = 97.36 Min = 95.02				Max = 97.51 Min = 96.04				Max = 97.66 Min = 96.05				Max =97.37 Min = 95.46			

Table 3: The results of the system when training set is 50% and the test set is 50 % .

One rule				Two rules				Three rules				Four rules				Five rules			
train	test	%	Av	train	test	%	Av	train	test	%	Av	train	test	%	Av	train	test	%	Av
97.37	97.07	97.22	4	97.95	96.48	97.22	2.5	97.37	79.66	97.51	3.3	97.95	96.48	97.22	4	98.25	96.48	97.36	4
98.83	94.13	96.48	4	96.49	97.36	96.93	3	96.49	96.77	96.63	3.7	98.54	94.72	96.63	3.5	97.66	96.48	97.07	3
97.37	95.31	96.34	3	97.66	95.60	96.63	3.5	96.49	96.77	96.63	3.7	97.95	95.01	96.48	3	97.95	96.19	97.07	4.2
97.95	94.72	96.34	4	97.08	95.60	96.34	3.5	95.91	97.07	96.49	2.7	99.12	93.84	96.48	3.5	97.37	96.48	96.92	4.6
96.49	95.89	96.19	3	97.37	95.31	96.34	3.5	95.91	97.07	96.49	2.7	96.78	95.89	96.34	2.8	97.08	95.89	96.49	2.4
97.08	95.31	96.19	3	97.66	95.01	96.34	2.5	96.78	96.19	96.49	2.7	97.08	95.31	96.19	4	97.37	95.31	96.34	3
98.54	92.67	95.60	3	98.54	94.13	96.34	3	97.08	95.89	96.49	3.7	95.91	95.89	95.90	3	97.66	95.01	96.34	2.8
96.20	94.72	95.46	3	97.66	94.43	96.04	3.5	97.37	95.31	96.34	3.7	96.20	94.72	95.46	2.3	07.37	94.72	96.04	3
96.49	92.96	94.73	3	97.66	93.84	95.75	3	98.83	92.96	95.90	3.7	97.08	92.38	94.73	3	95.61	96.48	96.05	3.2
94.44	94.72	94.58	4	96.49	94.72	95.60	3.5	95.61	92.96	94.29	3	96.49	91.20	93.85	3	96.49	92.67	94.58	2.8
Max = 97.22 Min = 94.58				Max = 97.22 Min = 95.60				Max = 97.51 Min = 94.29				Max = 97.22 Min = 93.85				Max =97.36 Min = 94.58			

References

American Cancer Society Homepage 2008. Citing Internet sources available from : <http:// www.cancer.org>.

Goldberg DE, 1989. Genetic algorithm in search, optimization and machine learning. Reading, MA:Addison-Wesley.

Herrera F, Lozano M, Verdegay JL, , 1995. generating fuzzy rules from examples using genetic algorithms. In ; Bouchon-Meunier B, Yager RR, Zadeh LA, editors. Fuzzy genetic and soft computing . World Scientific.

Jain R, Abraham A, 2004. A comparative study of fuzzy classification methods on breast cancer data.

Lee CC, 1990. Fuzzy Logic in Control Systems: Fuzzy Logic Controller -Part 1. IEEE Transactions on Systems,Man. And Cybernetics, volume 20, no. 2 March/April.

Mangasarian OL, Setiono R, Goldberg W-H, 1990. pattern recognition via linear programming : Theory and application to medical diagnosis. In Coleman TF , Li Y, editors. Large-Scale Numerical Optimization.

Mendel JM, 1995. Fuzzy logic systems for engineering : a tutorial . Proceedings of the IEEE,83(3):345-377.

Merz CJ, Murphy PM, 1996. UCI repository of machine learning databases. http:::MLRepository.Html.

Michalewicz Z, 1996.Genetic Algorithms_Data structures_Evolution Programs, 3rd edition Heidelberg : Sppringer-Verlag.

Setiono R, 2000. Generating consise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence in medicine, (1893),205-217. doi:10.1016/s0933-3657(99)00041-X.

Van Leekwijck and Kerre, 1999. Defuzzification: criteria and classification. FuzzySets and Systems, volume 108, pp.159-178.

Yager RR, Filev Dp, 1994. Essential of Fuzzy Modeling and Control.Wiley. Zadeh LA. Fuzzy sets, 1965. Information and Control. vol 8, pp. 338-353.