A Fuzzy Rough Model Using Genetic Algorithm for Accuracy Enhancement

Mohamed S.S. Basyoni[1]* **Hesham A. Hefny[1]**

[1]Cairo University, Egypt
* Corresponding author's Email: m_basyoni2@ yahoo.com

Abstract: Fuzzy rules have been used extensively in data mining. Our model is proposed to construct Fuzzy Rules with considering criteria of accuracy. The model consists of five phases: (1) Automatic attributes fuzzification, (2) Feature selection via core and reduct, (3) Fuzzy rules extraction via SQL statement, (4) Calculates Accuracy and Confidence for each rule, (5) Genetic coding of fuzzy rules. The algorithm determines automatically the parameters width of each attribute then uses the rough set to reduce the number of attributes (Feature Selection) via core and reduct. The model then extracts fuzzy rules using SQL statements and calculates Accuracy and Confidence. Finally, our genetic model represents each fuzzy set by "Real number" to improve the accuracy. The model applied on Iris and Wine datasets and the result will be better than the other models in term of the number of fuzzy sets and classification rate for evaluating the accuracy of training and test instances.

Keywords: Rough set, Fuzzy logic, Core and reduct, SQL statements, Accuracy and confidence, Genetic algorithm.

1. Introduction

Rough set was proposed by Professor Pawlak [1, 2], a Polish mathematician in 1982, and was one of the mathematical tools for dealing with vagueness and uncertainty information. Rough set is oriented to discrete data, and discretization of continuous attributes will lead to information loss, which will greatly affect the quality of classification results. In real life datasets values of attributes could be both of symbolic and real-valued. Therefore, fuzzy logic used to solve the problem of hesitation through non-membership degree, which overcomes the problem of rough set by combining fuzzy logic with rough set together [3, 4, 5]. Another problem of traditional Fuzzy-Based Rough Model [6, 7] is that a user defines parameters of membership functions of fuzzy sets for each numerical attribute which will different from one user to another. To solve this problem, a new Automatic Fuzzy Rough Model proposed that can define those parameters automatically. The user defines only the number of fuzzy sets "n" of each attributes, and then the max and min values of each attribute are computed and calculate the width (Δ) that divides the universe of discourse of each attribute into intervals according to the number of fuzzy sets and

calculates the parameters width (δi) according to thewidth (Δ). Another strong drawback of the traditional rough set theory [8, 9] is the inefficiency of computing the core attributes and reduct and identifying the dispensable attributes, which limits the suitability of the traditional rough set model in data mining applications. To solve this problem, a rough sets model based on database systems has been introduced for this problem to redefine a core and reducts by using relational algebra such as Cardinality and Projection. Most traditional methods [10-13] used a heuristic method to generate a linguistic rule which generates a large number of rules and need other technique like (Genetic Algorithm – Based Rule Selection) to select a small number of significant rules to reduce the large number of candidate fuzzy rules. For this problem we use the algorithm for extracting fuzzy rules using SQL statements which generates efficient smaller number of fuzzy rules immediately without needing to run a genetic algorithm approach to do this step. After that, calculates automatically the Accuracy and Confidence of each rule then get the Accuracy and Confidence of all rules.

Multiobjective evolutionary algorithms (EAs) that use nondominated sorting and sharing [14, 15] have been criticized mainly for their: 1) O(MN)

computational complexity (where M is the number of objectives and

N is the population size ; 2) the need for specifying a sharing parameter but in our genetic model each fuzzy set represented by "Real number" from 0 to 9 forming a gene on chromosome (individual). Our algorithm using Genetic algorithm technique to generate another new fuzzy rule from the initial rules then calculates their accuracy again which will be greater than the old rules before using genetic algorithm. The proposed model is applied on the Iris and Wine datasets and the results compared with other models: Preselection with niches [16], NSGA-II (Non dominated sorting genetic algorithm II) [17, 18], ENORA (Evolutionary Non dominated sorting with Radial slots) [19, 20], AP- NSGA-II (Average-Point dominated sorting genetic algorithm II) [21], FRFS (Fuzzy Rough Feature Selection) [22], T-FRFS (Threshold Fuzzy Rough Feature Selection) [23] and C-FRFS (C-Means Fuzzy Rough Feature Selection) [24] algorithms in term of number of fuzzy sets (L) and classification rate for evaluating the accuracy of training and test instances (CR) to show its validity.

This paper is organized as follows: Basic concept of the rough set, Fuzzy logic and genetic algorithm in Section 2. In the same section we propose a new automated Fuzzy Based algorithm and the Rough Set Model Based on Database Systems technique is used to reduce the number of attributes. In section 3, we show the Proposed Linguistic Model that contains five phases on dataset in (X.Hu, T.Lin, and J.Han.2004) [25]. In section 4, shows the experiments and the results obtained for the problem of classification of the Iris data set and compare the result of our model with other models: Preselection with niches, ENORA, NSGA-II, FRFS, T-FRFS and C-FRFS. Finally, show the Contribution and conclusion in Section 5 and 6.

2. Basic concepts

2.1. Rough set and feature selection via core and reduct

Rough set theory was developed by Pawlak in the early 1980's [1, 2] and was one of the mathematical tools for dealing with vagueness and uncertainty information. A rough sets model based on database systems has been proposed to redefine a core and reducts by using relational algebra such as Cardinality (*Card*) to denote the *Count*, and Π for *Projection* operation.

Example 2.1

A collection of 8 cars (C1,C2,…C8) as in Table 1, with information about the attributes Weight (*W*), Door

Table 1. Cars table with attributes *W, D, S, C and M*

C_i	W	D	S	C	M
C_1	small	2	comp	4	large
C_2	small	4	sub	6	small
C_3	medium	4	comp	4	large
C_4	large	2	comp	6	small
C_5	large	4	comp	4	small
C_6	small	4	comp	4	large
C_7	large	4	Sub	6	small
C_8	small	2	sub	6	small

(*D*), Size (*S*) and Cylinder (*C*) are the condition attributes and Mileage (*M*) is the decision attribute.

Definition 2.1.1.
[25] An attribute $Aj \in A$ is a *core* if:

$$Card\ (\Pi\ (A - Aj + D)) \neq Card\ (\Pi\ (A - Aj))$$

Where:
Card: *Cardinality* to define the *count* of attribute
Π: *Projection* operation   *A:* Condition attributes
D: Decision attribute
For example, in Table 1

$$Card\ (\Pi\ (A - Aj + D)) = Card\ (\Pi\ (D, S, C, M) = 6,$$
$$Card\ (\Pi\ (A - Aj)) = Card\ (\Pi\ (D, S, C) = 5$$

Therefore, from Table 1, the attribute *Aj=W* is a core.
Definition 2.1.2.
[25] An attribute $Aj \in A$ is a *dispensable* if:

$$Card\ (\Pi\ (A - Aj + D)) = Card\ (\Pi\ (A - Aj)).$$

For example, in Table 1

$$Card\ (\Pi\ (A - Aj + D)) = Card\ (\Pi\ (W, S, C, M) = 6,$$
$$Card\ (\Pi\ (A - Aj)) = Card\ (\Pi\ (W, S, C) = 6$$

Thus, from Table 1, the attribute *Aj=D* is a dispensable.
Definition 2.1.3.
[25] The *mer* value of an attribute *Aj* in *A* is defined as:

$$Mer\ (Aj, A, D) = 1 - (Card\ (\Pi\ (A - Aj + D)) / Card\ (\Pi\ (A - Aj)))$$

Mer (*Aj, A, D*) reflects the degree of contribution made by the attribute *Aj* to the dependency only between *A* and *D*. For example, in Table 1

$$Card\ (\Pi\ (A - Aj + D)) = Card\ (\Pi\ (D, S, C, M) = 6,$$
$$Card\ (\Pi\ (A + D)) = Card\ (\Pi\ (W, D, S, C, M) = 8$$
$$Mer\ (W, D, W, S, C, M) = 1 - 6/8 = 0.25$$

Table 2. the eleven examines on comfortable in a car

p	Weight	Comfor
1	45	yes
2	50	yes
3	55	yes
4	60	no
5	65	yes
6	70	no
7	75	no
8	80	no
9	85	no
10	90	no
11	95	no

2.2. Fuzzy linguistic classification

Fuzzy logic has been applied very successfully in many areas where conventional model based approaches are difficult or not cost-effective to implement. However, as system complexity increases, reliable fuzzy rules and membership functions used to describe the system behavior are difficult to determine. The consequent of the fuzzy rule is the accuracy of the classification. The confidence of the association rules are used to choose the rules. The fuzzy rule is considered to be the associate rule, which is described as follows: Aq => Cq. Where, Aq is the antecedent of the rule, and Cq is the consequent of the rule. These two conceptions are extended to the fuzzy association rule. The forms proposed by Ishibuchi and Nakashima [26] are used in this paper to get the confidence of the association rules as:

$$C(Aq \rightarrow Cq) = \frac{|D(Aq) \cap D(Cq)|}{|D(Aq)|}$$
$$= \frac{\sum \mu Aq(x) \cap \mu Cq(x)}{\sum \mu Aq(x)} \qquad (1)$$

Where,
Aq: the antecedent part of the rule
Cq: the consequent part of the rule
$|D(Aq)| = \Sigma \mu Aq(x)$: the cardinality of a fuzzy set
$|D|$: no. of all patterns

Example .2.2
Suppose the eleven examinees if they comfortable in a car or not as shown in Table 2
Suppose the fuzzy set *medium* =

$$\{ \frac{0.33}{55}, \frac{0.67}{60}, \frac{1}{65}, \frac{1}{70}, \frac{1}{75}, \frac{0.67}{80}, \frac{0.33}{85} \}$$

Now we will calculate *C* (*medium => uncomf*) and *C* (*medium => comf*) as following:

The total compatibility grade with the fuzzy set *medium* is calculated as:

$| medium | = 0.33 + 0.67 + 1 + 1 + 1 + 0.67 + 0.33 = 5$

1- From the table, $| medium \cap uncomf | = 0.67 + 1 + 1 + 0.67 + 0.33 = 3.67$
The confidence of the association rule
"medium => uncomf" calculated as:

$C (medium => uncomf) = 3.67 / 5 = 0.73$

and
2- From the table, $| medium \cap comf | = 0.33 + 1 = 1.33$
The confidence of the association rule
"medium => comf" calculated as:

$C (medium => comf) = 1.33 / 5 = 0.27$

and

Since *C* (*medium => uncomf*) is larger than *C* (*medium => comf*),

The association rule: "*medium => uncomf*" is better than "*medium => comf*".
A fuzzy set represented by its membership function and defined by parameters over the domain called the universe of discourse

2.2.1. Triangular Membership Function

Triangular (X: a, b, c)
$$= \begin{cases} 0 & x \le a \\ (x-a)/(b-a) & a \le x \le b \\ (c-x)/(c-b) & b \le x \le c \\ 0 & x > c \end{cases}$$

For "n" fuzzy sets of each attribute that divided it into intervals with the width between center parameter b_i and b_{i+1} (Δ) where:

$$\Delta = \frac{Max - Min}{n - 1} \qquad (2)$$

With: Min = Min – R1,   Max = Max + R2, and
$\delta_i = (\Delta + 1) / 2$
We can also calculate the parameters a, b and c by the following equations:

$$A_i = (a_i, b_i, c_i)$$
$$b_i = Min + (i - 1) \times \Delta \qquad (3)$$

With $a_i = b_i - \delta_i$, $c_i = b_i + \delta_i$
Where:
Δ: the width between center b_i and b_{i+1},
Max: the maximum value of the attribute

Min: the minimum value of the attribute,
n: the number of fuzzy sets
R1: A value subtracted from min to make it integer value,
R2: the value added to max to make it integer value,
δ_i: the parameters width between bi and ai, ci

2.3 Genetic algorithm

A genetic algorithm [27, 28] is a way to perform a heuristic search in a solution space based on the evolutionary ideas of natural selection and genetics. It is an iterative procedure maintaining a population of structure of candidate solutions to specific domain, which each candidate solution is called a chromosome (or individual). Genetic algorithms have been widely proposed to generate fuzzy if-then rules and tune the membership function of fuzzy sets in fuzzy rules. The chromosome representation, selection, crossover, mutation, and fitness function computation are the key elements of GA. The procedure of GA is as follows. Populations of chromosomes are initialized randomly. The fitness of each chromosome is computed. Two chromosomes are selected from the population according to the fitness value. The single-point crossover operator with crossover probability is applied to produce an offspring. Thereafter, uniform mutation operator is applied on produced offspring with mutation probability to generate new offspring. The new offspring is placed in new population. The selection, crossover, and mutation operations will be repeated on current population until the new population is complete.

3.   The proposed linguistic model

3.1. Phase one: automatic attributes fuzzification

We use our automated Fuzzy-Based Rough Decision Model algorithm on dataset in [25].
Suppose we used "Triangular Membership function" and define "3" fuzzy sets: small, Medium and large. Fuzzifing numerical data of attribute D will be as the following:
Min = 2 and Max = 4 So, R1 = 1 and R2 = 0 then:
Min = 1 and Max = 4
The width between bi and bi+1 is:
Δ = 4-1/ (3-1) = 1.5,
δ_i = (1.5+1) / 2 = 1.25
The 3 parameters a,b and c of the three fuzzy sets F1,F2 and F3 will be as following:
1- F1 ⟶ small, δ_1 =1.25   b1 = a1 = 1, c1 = 2.25
2- F2 ⟶ Medium, δ_2 =1.25
b2 = 2.5, a2 = 1.25, c2 = 3.75
3- F3 ⟶ large, δ_3 = 1.25   b3 = c3 = 4, a3 = 2.75.
The   result   will   be   shown   as   Fig   .1
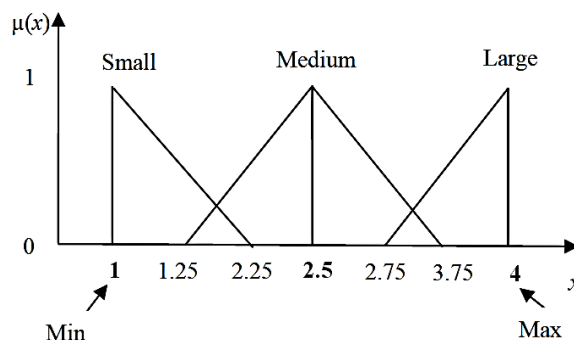

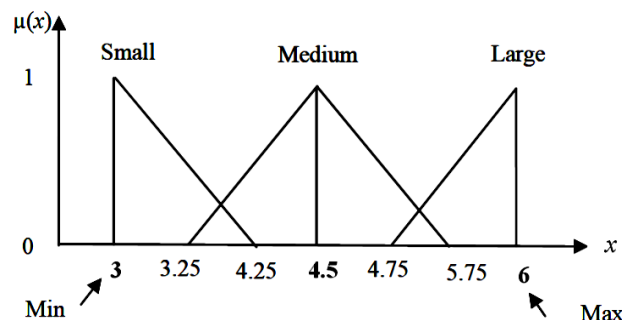Figure. 1 Door attribute

Figure. 2 Cylinder attribute

Table 3. Table of Cars with attributes *D and C*

Ci	W	D	S	C	M
C1	small	medium	comp	medium	large
C2	small	high	sub	high	small
C3	medium	high	comp	medium	large
C4	large	medium	comp	high	small
C5	large	high	comp	medium	small
C6	small	high	comp	medium	large
C7	large	high	Sub	high	small
C8	small	medium	sub	high	small

Suppose we also used "Triangular Membership function" and define "3" fuzzy sets: small, Medium and large.
Fuzzifing numerical data of attribute C will be as the following:
Min = 4 and Max = 6 So, R1 = 1 and R2 = 0 then:
Min = 3 and Max = 6
The width between bi and bi+1 is:Δ = 6-3/ (3-1) = 1.5,
δ_i = (1.5+1) / 2 = 1.25
The 3 parameters a,b and c of the three fuzzy sets F1,F2 and F3 will be as following:
1- F1 ⟶Small, δ_1 =1.2
b1 = a1 = 3, c1 = 4.25
2- F2 ⟶ Medium, δ_2 =1.25
b2 = 4.5, a2 = 3.25, c2 = 5.75
3- F3 ⟶ large, δ_3 = 1.25   b3 = c3 = 6, a3 = 4.75.
The result will be shown as Fig. 2
The result of fuzzifying numerical data of the two attributes D and C of table 1 will be as in Table 3.

572

Algorithm 1: All Core attributes

Input: Table T (A, D)
Output: Core attribute of table T.
Method:
1. Set Core = 0
2. For each attribute $A_j \in A$
{
If Card (Π (A – A_j + D)) ≠ Card (Π (A – A_j))
Then Core = Core $\cup$ A_j
}

Algorithm 2: Computing reduct

Input: Table T (A, D)
Output: Reduct (Red, D)
Method:
1. Run Alg. 1 to get a core attr. of table T
2. Red = T
3. R = A - Red
4. Compute mer values for all attributes of R
5. Sort attributes in R based on mer values in decreasing order
6. Choose an attribute A_j with the largest mer values
7. Red = Red $\cup$ {A_j}, R = R - {A_j}
8. If K (Red, D) = 1, then terminate, otherwise go back to Step 4

3.2. Phase two: feature selection via core and reduct

The proposed algorithm for finding all core attributes based on the relational database system without calculation of the lower and upper approximations is described in Algorithm 1 [25]

Algorithm 2 computes a minimal attribute subset (reduct) [25]

The output of this algorithm is a reduct *W*, *D* and *S* from the data in Table 3.

3.3. Phase three: fuzzy rules extraction via SQL statement

Most traditional methods [10-13] used a heuristic method to generate a linguistic rule for each cell of the pattern space which generates a large number of rules and need another algorithm like a genetic algorithm approach (Genetic Algorithm – Based Rule Selection) to select a small number of significant rules from them to reduce those large number of candidate fuzzy rules. This means that if we have *K* fuzzy sets for each of *n* attributes, then the number of linguistic rule = K^n rules. For example, if we have 3 fuzzy sets and 4 attributes then the number of fuzzy rules will be (81) rules which will be a big number of rules and need genetic algorithm to reduce them. For this problem a new algorithm for extracting fuzzy rules using SQL statements is proposed that generates efficient minimum number of fuzzy rules immediately without needing to run any other approach to do this step. The

W is 'Large' And D is 'large' And S is' comp' ==> M is Small
W is 'Large' And D is 'large' And S is 'sub' ==> M is Small
W is 'Small' And D is 'large' And S is 'comp' ==> M is Large
W is 'Small' And D is 'large' And S is 'sub' ==> M is Small
W is 'Small' And D is 'Medium' And S is 'comp' ==> M is Large

Figure. 3 The initial fuzzy rough rules

proposed algorithm for extracting fuzzy rules using SQL statements running by 3 steps:
1- Create temp table contains the selected reduct and decision attribute
2- Get all equivalence classes for the selected reduct
3- Get all decision rules for the current equivalence class

We input a number of instances from table as training instances then the algorithm *randomly* generates the fuzzy rules from them. Therfore, the remaining instances will be the test records. After that, the algorithm calculates the accuracy of those fuzzy rules. In Table 3 after reducing attributes (Feature Selection) and generating a list of Reducts we can easily select *Mileage* attribute as a decision attribute then remaining attributes *W*, *D* and *S* will be a condition attributes. We randomly choose 5 instances as the training instances and 3 instances as the testing instances. Then the algorithm generates 5 fuzzy rules from those 5 records as in Fig. 3.

3.4. Phase four: calculates accuracy and confidence for each rule

For the fuzzy rules in Fig. 3 that generated, the algorithm computes the confidence and the accuracy of each fuzzy rough rule, and then it calculates the total accuracy of all rules which will be 0.63.

3.5. Phase five: genetic coding of fuzzy rules

The Genetic algorithm is applied on the pervious fuzzy rules. In our genetic model each fuzzy set represented by "Real number" from 0 to 9. In addition to the linguistic terms, "don't care" is also used for each attribute as an additional antecedent fuzzy set.
For example we represent three fuzzy sets Small, Intermediate and Large as:

1: Small, 2: Intermediate, 3: Large, 0: Do not care
Therefore, if we have the following rule of five attributes and each attribute have the three previous fuzzy sets as:

Rule *Ri:* If x1 is Large and x2 is Intermediate and x3 is Small And x5 is Large then Class *Cj*
Therefore, the representation of this rule as chromosome will be as Fig. 4.

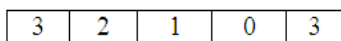The consequent class *Cj* of each fuzzy if-then rule is determined by the test patterns in the fuzzy subspace

3	2	1	0	3

Figure. 4 Representation of rule as chromosome

Algorithm 3: The Proposed Genetic-based Algorithm for generate the Fuzzy Classification Rules

Step1: Parameter Specification. Population size, maximum generation MAXGEN, and a fitness value condition.

Step2: Initialization. Specify an initial population, the number of antecedent of fuzzy sets and their representation as a strings of length n, and set generation gen = 1.

Step3: Genetic Operations. Calculate the fitness value of each linguistic rule in the current population using the test patterns.

Generate new linguistic rule using selection, crossover, and mutation from existing linguistic rule in the current population.

Step4: Generation Update. Remove the worst linguistic rules from the current population and add the newly generated rules to the current population gen = gen +1.

Step5: Termination Test. If gen = MXAGEN or get a fitness value then go to Step 6 else go to Step3.

Step6: Best Rules. Take the individual with maximum fitness value as the optimal fuzzy rule.

---

specified by the antecedent fuzzy sets. After that, set parameters, max number of generations MAXGEN, population size and a fitness value condition. We set the max number of generations MAXGEN is 100, the accuracy value condition we need to reach is 98% and population size is 20. Also the Crossover probability pc is 0.8; the Mutation probability pm is 1/ (string length). Next the fitness value of each linguistic rule in the current population is evaluated. Let S be the set of fuzzy rules in the current population. The evaluation of fitness value of each fuzzy rule is performed by classifying all given test patterns by the rule set S using the single winner-based method. The rule receives a unit reward when it correctly classifies test patterns. After all the given test training patterns are classified by the rule set S, the fitness value $fitness(Rq)$ of each linguistic rule Rq is calculated as the following:

$$fitness(Rq) = NCP(Rq) \qquad (4)$$

Where, $NCP(Rq)$ is the number of correctly classified test patterns by Rq.

After that, the fitness value $NCP(S)$ of the rule set S is calculated as the following:

$$NCP(S) = \sum NCP(Rq) \qquad (5)$$

Where
$Rq \in S$

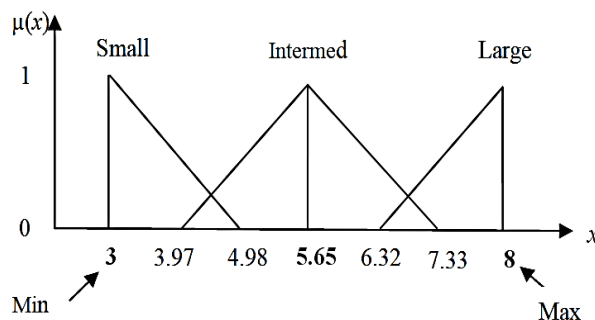

Figure. 5 Sepal length *(SL)* attribute

We propose our constrained genetic model, with two objectives:

Maximize NCP(S) and *Minimize | S |*

Where, *NCP(S)* is the number of correctly classified test patterns by the rule set *S*, and *| S |* is the number of fuzzy if-then rules in *S*. The Proposed Genetic-based Algorithm for generation the Fuzzy Classification Rules is shown in Algorithm 3.

When we run the genetic algorithm on the initial fuzzy rules in Fig 3, the algorithm generates other new Eq. (2) rules and adds them to the initial fuzzy rules. The accuracy of the new 7 rules from randomly 5 rows will computed to be 0.94 instead of 0.63

4. Experiments and results

The Iris data contains 150 measurements of four features (sepal length *(SL)*, sepal width *(SW)*, petal length (*PL*), petal width (*PW*)) from each of three species Class *(RS)* (Setosa, Versicolor, Virginica). Therefore, there are four independent variables that form the antecedent of the rule, and three dependent variables that form the consequence of the rule. From each species, there are 50 observations regarding sepal length, sepal width, petal length and petal width (in cm). We choose "Triangular Membership function", then define "3" fuzzy sets: small, Intermediate and large. Fuzzifing numerical data of attribute *(SL)* will be as the following:

Min = 4.3 and Max = 7.9 So, R1 = 1 and R2 = 0.1
So, Min = 3.3 and Max = 8
Δ = 8-3.3/ (3-1) = 2.35
δi = (2.35+1) / 2 = 1.68, therefore:
1- A1 ⟶ small, $\delta 1$ =1.68
b1 = a1 = 3.3, c1 = 4.98
2- A2 ⟶ Intermediate, $\delta 2$ =1.68
b2 = 5.65, a2 = 3.97, c2 = 7.33
3- A3 ⟶ large, $\delta 3$ = 1.68
b3 = c3 = 8, a3 = 6.32.
The result will be shown as Fig .5.

Fuzzifing numerical data of attribute *(SW)* will be as the following:
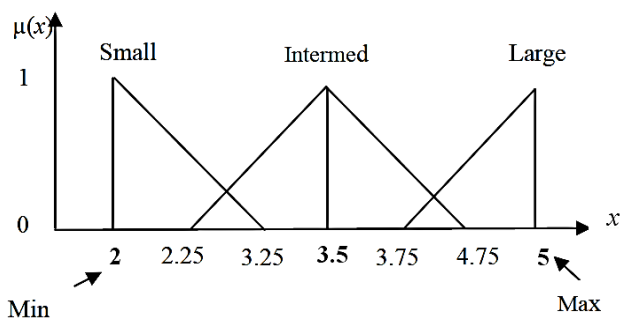Min = 2.2 and Max = 4.4 So, R1 = 0.2 and R2 = 0.6

Figure. 6 Sepal width *(SW)* attribute

Table 4. Iris data table after fuzzifing condition attributes

no	SL	SW	PL	PW	RS
1	Intermed	Intermed	small	small	setos
2	Intermed	Intermed	small	small	setos
3	Intermed	Intermed	small	small	setos
4	Intermed	Intermed	small	small	setos
⋮	⋮	⋮	⋮	⋮	⋮
51	Large	Intermed	Intermed	Intermed	versicol
52	Intermed	Intermed	Intermed	Intermed	versicol
53	Large	Intermed	Intermed	Intermed	versicol
⋮	⋮	⋮	⋮	⋮	⋮
101	Intermed	Intermed	Large	Large	vergin
102	Intermed	small	Intermed	Intermed	vergin
103	Large	Intermed	Large	Intermed	vergin
⋮	⋮	⋮	⋮	⋮	⋮
150	Intermed	Intermed	Intermed	Intermed	vergin

So,   Min = 2 and Max = 5

$\Delta = 5-2 / (3-1) = 1.5$

$\delta i = (1.5+1) / 2 = 1.25$, therefore:

1- A1 ⟶ small, $\delta 1 = 1.25$

$b1 = a1 = 2, c1 = 3.25$

2- A2 ⟶ Intermediate, $\delta 2 = 1.25$

$b2 = 3.5, a2 = 2.25, c2 = 4.75$

3- A3 ⟶ large, $\delta 3 = 1.25$

$b3 = c3 = 5, a3 = 3.75.$

The result will be shown as Figure .6

In the same manner fuzzifying numerical attributes *(PL)* and *(PW).* The result will be as shown in Table 4.

We can easily select *Class (RS)* attribute as a

Figure. 7 The initial fuzzy rules for iris data before running Genetic algorithm

Table 5. The *Confid* and *Accuracy* of each Fuzzy Rule before Genetic algorithm

Fuzzy Rules No.	frequency of rows	Confidence	Accuracy
R1	1	0.82	2
R2	5	0.71	4
R3	2	0.66	0
R4	3	0.84	1
R5	2	0.83	0
R6	1	0.63	1
R7	19	0.58	2
R8	5	0.71	1
R9	4	0.6	5
R10	28	0.7	15
R11	1	0.66	2
R12	29	0.55	1
Total	*100*		*34*
CR-training	*1.00000*		
Accuracy % (CR – evaluation)			*68%*

decision attribute in Table 4 then the remaining attributes *SL, SW, PL* and *PW* will be a four condition attributes. Let the number of fuzzy sets is (*L*), the number of fuzzy rules is (*M*) and the classification rate (*CR*) for evaluating the accuracy of training and test instances. We randomly choose 100 instances as the training instances and 50 instances as the testing instances. The algorithm generates 12 fuzzy rules from those 100 records using triangular membership functions as Fig .7.

The training classification rate (*CR*-training) of those

Figure. 8 The fuzzy rules for iris data after running genetic algorithm

Table 6. The *Confid* and *Accuracy* of each Fuzzy Rule after Genetic algorithm

Fuzzy Rules No.	frequency of rows	Confidence	Accuracy
R1	1	0.82	2
R2	**5**	**0.76**	**5**
R3	2	0.66	0
R4	3	0.84	1
R5	2	0.83	0
R6	1	0.63	1
R7	**19**	**0.85**	**10**
R8	5	0.71	1
R9	4	0.6	5
R10	28	0.7	15
R11	1	0.66	2
R12	**29**	**0.75**	**6**
Total	*100*		*48*
CR-training	*1.00000*		
Accuracy % (CR – evaluation)			*96%*

rules will be 1.00000 and the number of fuzzy sets L is 3 fuzzy sets. For the initial fuzzy rules in figure 8 that generated, The confidence and the accuracy of each fuzzy rough rule will calculated, the confidence computed using the Average operator that takes the average confidence value of the fuzzy sets of rule then compute the total accuracy of all rules as in the Table 5.

The final step is running the Genetic algorithm on

Table 7. The best results obtained for Iris data

Models names	M	L	CR-training	CR-evaluation
Preselection w. niches [16]	10	9	0.989899	94
NSGA-II [17, 18]	10	9	0.989899	94%
ENORA [19, 20]	10	8	0.989899	92%
	13	8	1.000000	94%
AP- NSGA-II [21]	12	5	1.000000	95%
Model of This paper	**12**	**3**	**1.000000**	**96%**

those fuzzy rules. The fuzzy sets in each rule represented in a chromosome as a number from 0 to 9. We represent each fuzzy set by real number as following:

1: Small, 2: Intermediate, 3: Large, 0: Do not care

Finally, set parameters, max number of generations MAXGEN, population size and a fitness value condition. We set the max number of generations MAXGEN is 100, the accuracy value condition we need to reach is 98% and population size is 20. Also the crossover probability is 0.8. Then the algorithm generates other new rules and adds them to the fuzzy rules. The algorithm adds Eq. (3) new rules that shown in Fig. 8 in bolded font. The accuracy of those new 12 rules from randomly 100 rows will computed to be 96% instead of 68%. The results will be as following in the table 6 and Fig. 8.

We propose our constrained optimization model, with two objectives:

Maximize CR and *Minimize M* and *L* where, L: The number of fuzzy sets, M: The number of fuzzy rules, CR: The classification rate for evaluating the accuracy of training and test instances.

Finally, we compare our model with Preselection with niches [16], NSGA-II (Non dominated sorting genetic algorithm II) [17, 18], ENORA (Evolutionary Non dominated sorting with Radial slots) [19, 20] and AP- NSGA-II (Average-Point dominated sorting genetic algorithm II) [21] algorithms on iris dataset. The result shows that our algorithm is better than the others in term of Accuracy (CR –evaluation) and CR – training and also in the number of fuzzy sets. The chosen solution is bolded as shown in Table 7.

Another example of the Wine dataset (178 records), with information about 13 attributes as a condition attributes and *Classes* as a decision attribute which contain the following three classes:

Class 1 has 59 records   Class 2 has 71 records
Class 3 has 48 records

We also choose 100 instances as the training instances and 78 instances as the testing instances.

Figure. 9 The initial fuzzy rules for wine data before running Genetic algorithm

Figure. 10 The fuzzy rules for wine data after running Genetic algorithm

Table 8. The best results obtained for iris and wine data

Dataset	No. of rows	Reduct Size / Accuracy						
		FRFS [22]		T-FRFS [23]		C-FRFS [24]		Our Model	
Iris	150	4	94.82	3	94.52	3	94.52	**4**	**96**
Wine	178	5	95.38	4	94.82	4	95.51	**5**	**97.6**

Then the algorithm generates reduct size of only 5 attributes from 13 condition attributes and generates 13 fuzzy rules from those 100 records using triangular membership functions as Fig. 9.

The training classification rate (*CR*-training) of those rules will be 1.00000 and the number of fuzzy sets *L* is also 3 fuzzy sets. For the initial fuzzy rules in Fig. 9 that generated, the algorithm will automatically calculate the accuracy of rules before running Genetic algorithm which will be 0.74. The algorithm after using Genetic algorithm generates other new rules and adds them to the fuzzy rules. The algorithm adds Eq. (4) new rules that shown in Fig. 10 in bolded font. The accuracy of new 13 rules from randomly 100 rows will computed to be 97.6% instead of 74% as Fig. 10.

Also we compare our model with FRFS (Fuzzy Rough Feature Selection) [22], T-FRFS (Threshold Fuzzy Rough Feature Selection) [23] and C-FRFS (C-Means Fuzzy Rough Feature Selection) [24] algorithms on iris and Wine UCI machine learning repository datasets in term of Accuracy and reduct size. The result shows that our algorithm is better than the others in term of Accuracy with this reduct size. The chosen solution is bolded as shown in Table 8.

5. Contribution

First, before using our automated Fuzzy Rough algorithm on iris dataset and Wine dataset, the user should defines parameters of membership functions of fuzzy sets for each numerical attribute from his view which is different from one user to another, but by using the proposed automated algorithm the parameters defined automatically after choosing only the number of fuzzy sets.

Second, after using our automated Genetic-Based Fuzzy Rough Model on the initial fuzzy rules that generated from iris dataset as Fig. 7, the algorithm adds Eq. (3) new rules that shown in Fig 8 in bolded font with only 3 fuzzy sets (L) and with accuracy (CR – evaluation) 0.96% as shown in Table 6 and with also training classification rate (CR –training) 1.000000. This result will be better than the other models (Preselection with niches, ENORA, NSGA-II and AP-NSGA-II) in term of number of fuzzy sets (L) and Accuracy (CR –evaluation) and test instances (CR – training) as shown in Table 7.

In the same manner on initial fuzzy rules that generated from iris and Wine dataset as Figs 7 and 9 comparing with the other models (FRFS, T-FRFS and C-FRFS). In the wine dataset, after using the rough set algorithm via Feature selection to compute the reduct then the reduct size will be only 5 attributes from 13 attributes. The genetic algorithm also adds Eq. (4) new rules that shown in Fig 10 in bolded font with accuracy 97.6% for wine dataset and with accuracy 96% for iris dataset as the above result. This result is better than those models in term of accuracy with this reduct size as shown in Table 8.

6.   Conclusion

We have presented a flexible method for handling the Iris data and Wine dataset classification problem. A Genetic-Based Fuzzy Decision Model is proposed to be more flexible than the Traditional Genetic fuzzy models. First in traditional fuzzy rough models a user defines the parameters of membership functions of fuzzy sets for each numerical attribute which will different from one user to another but our algorithm can define them automatically by determine the max and min values and the number of fuzzy sets "n" then it finds the width (Δ) that divides the universe of discourse of each attribute into intervals according to the number of fuzzy sets and calculates the parameters width (δi) according to the width (Δ). After Fuzzifing original information system a Rough Sets Model Based on Database Systems has been applied to redefine some concepts such as core attributes and reducts by using relational algebra. Most traditional methods used a heuristic method to generate a linguistic rule which generates a large number of rules and need a genetic algorithm approach to select a small number of significant rules from them (Genetic Algorithm – Based Rule Selection) to reduce those large number of candidate fuzzy rules. For this problem we use the algorithm for extracting fuzzy rules using SQL statements which generates efficient smaller number of fuzzy rules in terms of accuracy immediately without needing to run a genetic algorithm approach to do this step. After that, our model calculates confidence and accuracy of each rule. Multiobjective evolutionary algorithms (EAs) that use nondominated sorting and sharing have been criticized mainly for computational complexity and needing for specifying a sharing parameter but in our genetic model each fuzzy set represented by "Real number" from 0 to 9 forming a gene on chromosome (individual). Our algorithm using Genetic algorithm technique to generate another new fuzzy rule from the initial rules then calculates their accuracy again which will be greater than the old rules. The proposed model is applied on the Iris and Wine datasets and the results compared with other models:

Preselection with niches, NSGA-II (Non dominated sorting genetic algorithm II), ENORA (Evolutionary Non dominated sorting with Radial slots), AP- NSGA-II (Average-Point dominated sorting genetic algorithm II), FRFS (Fuzzy Rough Feature Selection), T-FRFS (Threshold Fuzzy Rough Feature Selection) and C-FRFS (C-Means Fuzzy Rough Feature Selection) algorithms. The result shows that our algorithm is better than the other models in term of Accuracy (CR – evaluation) and in CR –training data and also in the number of fuzzy sets.

Conflicts of interest

I declare no conflict of interest. I have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

I declare that this work is original and has not been published before and is not currently being considered for publication elsewhere.

Author contributions

Conceptualization, Mohamed S.S. Basyoni carried out methodology, software, validation, formal analysis, investigation, resources, writing original draft preparation, writing review and editing.

References

[1]   T. K. Sheeja and A. S. Kuriakose, "A novel feature selection method using fuzzy rough sets", *Comput. Ind*, 97, 111–116, 2018.

[2]   Y. Li, S. Wu, Y. Lin, and J. Liu, "Different classes' ratio fuzzy rough set based robust feature selection", *Knowledge-Based Systems*, Vol. 120, pp. 74–86, 2017.

[3]   H, Soleimani. K, Govindan, H, Saghafi, and H, Jafari, "Fuzzy multi-objective sustainable and green closed loop supply chain network design", *Comput Ind Eng*, 109:191–203, 2017.

[4]   Q. Shen and A. Chouchoulas, "A rough fuzzy approach for generating classification rules", *Pattern Recognition 35*, 2425–2438, 2002.

[5]   X. Jiuping and L. Zhao, "A multi-objective decision-making with fuzzy rough coefficients to the inventory problem", *Information Sciences*, Vol. 180, pp. 679-696, 2010.

[6]   H. A. Qiang, W. Congxin, D. Chen, and S. Zhao, "Fuzzy rough set based attribute reduction for information systems with fuzzy decisions", *Knowledge-based systems,* Vol. 24 pp. 689-696, 2011.

[7]   Y. W. Li, Y. J. Lin, and J. H. Liu, "Feature selection for multilabel learning based on

kernelized fuzzy rough sets," *Neurocomputing*, Vol. 318, pp. 217–286, 2018.

[8] J. Dai, Q. Hu, H. Hu, and D. Huang, "Neighbor Inconsistent Pair Selection for Attribute Reduction by Rough Set Approach", *IEEE Trans. Fuzzy Syst*, Vol. 26, pp. 937–950, 2018.

[9] C. Wang, Y. Qi, and M. Shao., "A fitting model for feature selection with fuzzy rough sets", *IEEE Transactions on Fuzzy Systems*, Vol. 25, No. 4, pp. 741–753, 2017.

[10] M. T. M. Emmerich and A. H. Deutz, "A tutorial on multiobjective optimization", *Fundamentals and Evolutionary Method*s, Nat., Comput., Vol. 17, No. 3, pp. 585–609, 2018.

[11] A. K. Ghoshal, N. Das, S. Bhattacharjee, and G. Chakraborty, "A fast parallel genetic algorithm based approach for community detection in large networks", In: *Proc. of International Conference on Communication Systems & Networks (COMSNETS)*, Bengaluru, India, pp. 95–101, 2019.

[12] R. Guido and D. Conforti, "A hybrid genetic approach for solving an integrated multi-objective operating room planning and scheduling problem", *Comput Oper Res*, Vol. 87, pp. 270–28, 2017.

[13] T. Harada and E. Alba, "Parallel genetic algorithms: a useful survey", *ACM Computing Survey*, Vol. 53, No. 4, pp. 1–39, 2020.

[14] Y. A. Elnaga and S. Nasr, "Modified evolutionary algorithm and chaotic search for Bilevel programming problems", *Symmetry* Vol. 12, p. 767, 2020.

[15] L. Meenachi and S. Ramakrishnan, "Evolutionary sequential genetic search technique-based cancer classification using fuzzy rough nearest neighbour classifier", *Healthc Technol Lett* Vol. 5, pp. 130–135, 2018.

[16] F. Jimhez and A. E. G. Skanneta, "A Multi-objective Evolutionary Algorithm for Fuzzy Modeling", *IEEE*, 2001.

[17] K. Deb, A. Pratap, S.Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm NSGA-II", *Evolutionary Computation, IEEE Transactions on*, Vol. 6, No. 2, pp. 182-197, 2002.

[18] Y. Yusoff and M. Salihin, "Overview of NSGA-II for Optimizing Machining Process Parameters", *Published by Elsevier Ltd*, 2007.

[19] F. Jiménez, A. F. G. Skarmeta, G. Sánchez, and K. Deb, "An evolutionary algorithm for constrained multi-objective optimization", In: *Proc. of IEEE World Congress on Evolutionary Computation*, 2002.

[20] F. Jiménez and G. Sánchez, "Fuzzy Classification with Multi-objective Evolutionary Algorithms",

Springer-Verlag Berlin Heidelberg, pp. 730–738, 2008.

[21] P. Mohapatra and S. Roy, "AP-NSGA-II: An Evolutionary Multi-objective Optimization Algorithm Using Average-Point-Based NSGA-II", In: *Proc. of Fourth International Conference on Soft Computing for Problem Solving*, 2015.

[22] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection", *IEEE Trans. Fuzzy Syst* Vol. 17, p. 824, 2009.

[23] J. R. Anaraki and M. Eftekhari, "Improving fuzzy-rough quick reduct for feature selection", In *Proceedings of the 19th Iranian Conference on Electrical Engineering (ICEE)*, Tehran, Iran, 17–19 May 2011, pp. 1–6, 2011.

[24] R. Zhao, G. Lize, and X. Zhu, "Combining Fuzzy C-Means Clustering with Fuzzy Rough Feature Selection", *Beijing 100876*, China, 2019.

[25] X. Hu, T. Lin, and J. Han, "A New Rough Sets Model Based on Database Systems", *Fundamental Information 59*, No. 2-3 pp.135-152, 2004.

[26] H. Ishibuchi, T. Nakashima and M. Nii. "Classification and Modeling with Linguistic Information Granules", *Springer Berline Heidelberg New York*, 1998.

[27] M. Sikor, "Fuzzy Rules Generation Method for Classification Problems Using Rough Sets and Genetic Algorithms", *Springer-Verlag Berlin Heidelberg*, 2005.

[28] R. Rathi and D. P. Acharjya, "A rule based classification for vegetable production using rough set and genetic algorithm", *International Journal of Fuzzy System Applications (IJFSA)*, Vol. 7, No. 1, pp. 74–100, 2018.