IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A new hybrid Fuzzy Rough Model based on Genetic algorithm for data accuracy enhancement

Mohamed S.S.Basyoni, Ahmed M. Gadallah, Hesham A. Hefny
Phd student, Dr., prof. in Faculty of Graduate Studies for Statistical Research
Department of Computer and Information Sciences
Cairo University- Faculty of Graduate Studies for Statistical Research, Giza, Egypt

Abstract:

We live in the information era in which database size become massive due to the digital and technology revolution. Therefore, we introduce a new Enhanced hybrid algorithm which integrates the advantages of rough set theory and fuzzy set theory together with Genetic Algorithms (GAs). Our Model consists of four phases: (1) automatic attributes fuzzification, (2) Eliminate redundant attributes using rough set theory, (3) Generating Fuzzy rough rules then calculate automatically the accuracy and a fitness value (Confidence) for each rule, (4) Using the genetic algorithm for the Fuzzy rough rules to enhance their accuracy. In phase one, the user input the number of fuzzy sets of each attributes, the algorithm determine the maximum and minimum values of each attribute then calculates automatically the width (Δ) which divides the universe of discourse of each attribute into "n" intervals according to the number of fuzzy sets, after that calculates automatically the width (δ_i) according to the width (Δ). In phase two, the rough set techniques used to reduce the number of attributes that comes from phase one and produce fuzzy-rough rules. In phase three, the algorithm calculates the accuracy and the confidence (fitness value) of each fuzzy rough rule and calculates the total accuracy of all linguistic rules. In phase four, the genetic algorithm is running on the fuzzy-rough rules from phase three then it calculates the accuracy and the confidence of the new fuzzy rough rule again and calculates the total the accuracy of all rules. The accuracy of our algorithm that applied on Iris plants dataset before using our genetic algorithm from randomly 75 rows from 150 rows is 0.56 but after using our algorithm will be 0.95.

Index Terms **- Genetic algorithm, Fuzzy logic, Rough set, Accuracy, Automated Fuzzy - Based Rough Decision model.**

1. Introduction

Fuzzy rough decision models have been appeared in various recent researches due to its efficacy in generating potential decision rules especially in the cases of incomplete quantitative data sets. Genetic algorithms (GAs) have been proposed by various researches for tuning the rules obtained by such decision making models. However, the accuracy and simplicity of such models remains a research problem of such models that need further research.

Rough sets theory was developed by Pawlak in the early 1980's [16, 18, 25, 26] and has been applied successfully in a lot of domains. One of the major limitations of the traditional rough sets model is that it assumes that all attribute values are discrete. A real world data set always contains mixed types of data such as continuous valued, symbolic data, etc. Therefore all numerical or continuous data should convert to discretized data, here "Fuzzy Logic" can solve this problem to reduce information overload in a Fuzzy-Based Rough Decision Model [5, 7, 9, 11, 15, 19]. One drawback of traditional Fuzzy-Based Rough Decision Model is that the linguistic values (fuzzy sets) for numeric values of each attribute should determining by the membership functions of these linguistic terms which the user should define the parameters of those membership functions from his view which is different from one user to another. Therefore, we introduce a new automated Fuzzy-Based Rough Decision Model algorithm that can define those parameters automatically which h the user determine only the number of fuzzy sets then the algorithm automatically determine the maximum and minimum values of each attribute and calculates automatically the width (Δ) that divides the universe of discourse of each attribute into "n" intervals according to the number of fuzzy sets then the algorithm calculates automatically the width (δi) according to the width (Δ). Another drawback of the traditional rough sets model in the real applications is the inefficiency in the computation of core and reduct, because all the intensive computational operations are performed in flat files [1, 4, 14, 23]. In order to improve the efficiency of computing core attributes and reducts, a New Rough Sets Model Based on Database Systems has been introduced for this purpose [Hu, X., Lin, T., 2004], which redefine the core attributes and reducts

based on relational algebra to take advantages of the very efficient set-oriented database operations, such as *Cardinality* to denote the *count* and *Projection*.

The paper is organized as follows: We give an overview of the genetic algorithm in section 2, and give an overview of the rough set theory based on the model proposed by Pawlak [25, 26] with some examples, also give an overview of the fuzzy set theory. In section 3, we propose a new automated Fuzzy -Based Rough algorithm that can define the parameters of membership functions of linguistic values automatically. After that, we generate fuzzy rules. Finally, we use the genetic algorithm on the fuzzy rules that generate other efficient fuzzy rules in accuracy. In the same section, we explain the contributions of our model. Finally, we conclude with some discussions and our future works in Section 4 and section 5.

2. Basic Concepts

2.1. Genetic algorithm (GA)

Genetic algorithms are used as a machine learning tool for generating (evolving) a rule-based classification system [33]. A genetic algorithm randomly generates a pre-specified number of fuzzy rules [34, 35], which each rule set can be represented as a bit string. Thus a population of individuals corresponds to a single rule set. GA has a list of parameters that are included in the GA adaptation file as inputs and outputs parameters. The first input parameter of GA is fitness value, which be evaluated for any population of individuals. The fitness parameter assigned the value of the highest fitness value of any population member from the most recent generation. The GA uses two basic operations which are the mutation and the crossover. After implementing this stage it will yield a list of rules are generating [36, 37].

2.1.1. The Basic algorithm:

Initialize population Pi randomly
For i=1 to max generations
Evaluate fitness of individuals of population Pi
Repeat
  Select parents from Pi for reproduction
  Perform crossover on parents creating Pi+1
  Perform mutation of Pi+1
  Evaluate fitness of individuals of population Pi+1
Until best individual is good enough

2.1.2. Chromosome ("individual") representation

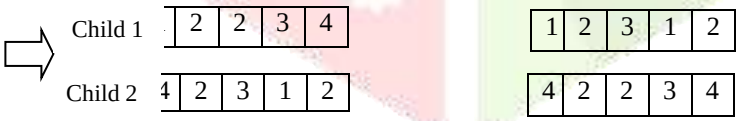
Each fuzzy set represented by integer number forming a gene on chromosome (individual) of fixed length that forms a population of fixed number of chromosomes

2.1.3. Selection Methods

The essential idea of a selection is that select the best individuals (candidate solutions) of the population to breed a new generation. The selection of the best individuals depends on their fitness

2.1.4. Crossover (recombination) operator

The crossover (or recombination) operator essentially swaps genetic material between two selected individuals from the population, called parent individuals



2.1.5. Mutation (bits flipped) operator

Mutation is an operator that acts on a single individual at a time. It usually applied with a small probability, typically much smaller than the crossover probability. Unlike crossover, which recombine genetic material between two or more parents, mutation replace (or flip) the value of a gene (a bit) with a randomly-generated value

Evaluate fitness of individuals of population, and then the above generational operators are repeated until a fitness condition has been reached.

2.2. Rough set theory (RST)

Rough set theory was developed by Pawlak in the early 1980's [8, 16, 18, 26] and has been successfully applied to knowledge acquisition as a powerful tool for data mining, decision analysis and forecast, knowledge discovery from database, and decision support system. The starting point of rough set is a dataset, which is usually organized into a table, and it is called information systems. In rough sets theory, the data is collected in a table, called decision table (or table in the database term) as shown in Table 1. We also assume that our data set is stored in a relational table with the form *Table (condition-attributes, decision-attributes)*. *C* is used to denote the condition attributes, *D* for decision attributes, where, $C \cap D = \Phi$. The main idea of rough set theory is based on the indiscernibility relation $[x]_R$ (equivalence classes).

Table 1: Cars table with attributes *Door, Size, Cylinder* and *Mileage*

Tuple -id	*Door*	*Size*	*Cylinder*	*Mileage*
t1	2	compact	4	high
t2	4	sub	6	low
t3	4	compact	4	high
t4	2	compact	6	low
t5	4	compact	4	low
t6	4	compact	4	high
t7	4	Sub	6	low
t8	2	sub	6	low

Definition 2.2.1 Let *T = (U, A)* be an information table with any set $B \subseteq A$ and $X \subseteq U$ We can approximate *X* using only the information contained in *B* by constructing the *R-lower* approximations of *X* that contain all objects in equivalence classes that "surely belong to X" denoted by where: $\underline{R}X$

$$\underline{R}X = \{x \mid [x]_R \subseteq X\}$$

Definition 2.2.2 Let *T = (U, A)* be an information table with any set $B \subseteq A$ and $X \subseteq U$ We can approximate *X* using only the information contained in *B* by constructing the *R-upper* approximations of *X* that contain all objects in equivalence classes that "may belong to X" denoted by where: $\overline{R}X$

$$\overline{R}X = \{x \mid [x]_R \cap X \neq \phi\}$$

Definition 2.2.3 The boundary area is the difference between the upper approximation and the lower approximation where:

$$BND\ (X) = \overline{R}X - \underline{R}X$$

The set X is said to be *rough (inexact)* if its *boundary region* is non-empty, otherwise the set is *crisp or exact* [16, 18]

Example 2.2.1

In table 1, as a dataset in (X.Hu, T.Lin, and J.Han.2004) [6] we have a collection of 8 cars (*t*1,*t*2,…*t*8) with information about the attributes *Door, Size, Cylinder* and *Mileage*, where *Door, Size* and *Cylinder* are the condition attributes and *Mileage* is the decision attribute. The attribute Tuple-id is just for explanation purpose and can be ignored. We can calculate the elementary sets, lower and upper approximations as follows:

The indiscernibility Relation of *C*:

IND (*R*) = *IND* ({*Door, Size, Cylinder*}) So, *R* = {*X1, X2, X3, X4, X5*} where,

[*X*]*R* = {{*t1*}, {*t2, t7*}, {*t3, t5, t6*}, {*t4*}, {*t8*}}

Let *D1* = {*x* | *Mileage* (*x*) = high}, then: *Y1* = {*t1, t3, t6*}

Lower approximation ={*X1*} = {*t1*}, *Upper* approximation = {*X1* ∪ *X3*} = {*t1, t3, t5, t6*}

Let *D2* = {*x* | *Mileage* (*x*) = low}, then: *Y2* = {*t2, t4, t5, t7, t8*}

Lower approximation = {*X2* ∪ *X4* ∪ X5} = {*t2, t4, t7, t8*}, *Upper* approximation = {*X2* ∪ *X2* ∪ *X3*} = {*t2, t3, t4, t5 t6, t7, t8*}

Boundary *C*/*D* = {*t3, t5, t6*}

This table is called "inconsistent table" that contains entities with the same condition attribute values but with different decision attribute values. Therefore, only 5 of 8 cars *t1, t2, t4, t7 t8* belong to the lower approximation of *D* based on *C*, while 3 of 8 cars fall in the boundary area. This fact indicates that the information of *Door, Size* and *Cylinder* collected so far is not consistent, for it is only good enough to make a classification model for the above five cars, but not enough to classify other three. In order to classify *t3, t5* and *t6*, more information is needed.

Example 2.2.2

In table 2, a new condition attribute, which is a *Weight* of cars, is added and collect information on for each car [6]. To classify the tuples in the boundary area, with the condition attributes set *C = {Weight, Door, Size, Cylinder}*, we can calculate the lower, upper approximations and boundary area as below:

Table 2: Cars table with attributes *Weight, Door, Size, Cylinder* and *Mileage*

Tuple -id	*Weight*	*Door*	*Size*	*Cylinder*	*Mileage*
t1	low	2	compact	4	high
t2	low	4	sub	6	low
t3	medium	4	compact	4	high
t4	high	2	compact	6	low
t5	high	4	compact	4	low
t6	low	4	compact	4	high
t7	high	4	Sub	6	low
t8	low	2	sub	6	low

The indiscernibility Relation of *C*:

IND (*R*) = *IND* ({*Weight, Door, Size, Cylinder*}), *R* = {*X1, X2, X3, X4, X5, X6, X7, X8*} where,

[*X*]*R* = {{*t1*}, {*t2*}, {*t3*}, {*t4*}, {*t5*}, {*t6*}, {*t7*}, {*t8*}}

Let *D1* = {*x* | *Mileage* (*x*) = high}, then: *Y1* = {*t1, t3, t6*}

Lower approximation = {*t1, t3, t6*},

Upper approximation = {*t1, t3, t6*}

Let *D2* = {*x* | *Mileage* (*x*) = low}, then: *Y2* = {*t2, t4, t5, t7, t8*}

Lower approximation = {*t2, t4, t5, t7, t8*},

Upper approximation = {*t2, t4, t5, t7, t8*}

Boundary *C*/*D* = ϕ

From above, one can see, this table is "consistent table".

2.3. Fuzzy Logic (FL)

Fuzzy logic was developed by Lotfi A. Zadah in the 1960's [15, 17, 19, 30], which is an extension of conventional Boolean logic constructed to deal with imprecise information i.e., statements that are neither completely true nor completely false. Fuzzy logic works through the use of fuzzy sets (FS). A fuzzy set is a set whose boundaries are not clearly defined. We can define a subset A of a given universal set U (the universe of discourse) as containing all elements x of U for which the element x is member of a fuzzy set A. Each

element belongs to a set A to a certain degree, called the degree of membership μA(x). Typically represented by a real-valued number in the interval [0...1]. This is contrast with conventional, crisp sets, whose boundaries are clearly defined. Each element either belongs or does not belong to a crisp set. So in term of degree of membership, we can say that an object can belong to a crisp set with only one of two possible degree of membership: 0 or 1. Fuzzy logic can provide a great way to convert numerical and real-valued data into categorical data by using linguistic values for numeric values and determining of membership functions of these linguistic terms.

Linguistic Variables (terms) can be defined over the domain of each attribute based on characteristics of that domain which take words as values like "hot, cold, small, medium, etc"

Each Linguistic Variable has a range of states called linguistic values, each of which is a fuzzy (linguistic) set defined over the same domain represented by its membership function this domain is called the *universe of discourse* of that linguistic variable.

3. Proposed model

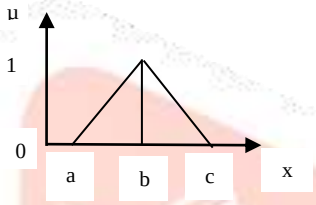
3.1. Phase 1: Automated Attributes Fuzzification

Fuzzy logic can provide a great way to convert numerical and real-valued data into categorical data by using linguistic values for numeric values and determining of membership functions of these linguistic terms which, every element in the universe of discourse is a member of the fuzzy set with some grade (degree of membership functions)

The traditional Fuzzy-Based Rough Decision Model has a major limitation that linguistic values (fuzzy sets) for numeric values of each attribute should determining by the membership functions of these linguistic terms which the user should define the parameters of membership functions of these linguistic values from his view which is different from one user to another. Therefore, we propose a new automated Fuzzy-Based Rough Decision Model algorithm which can define the parameters of membership functions of these linguistic values automatically that the user determine only the number of fuzzy sets (linguistic values) then the maximum and minimum values of each attribute are determined automatically then the algorithm calculates the width (Δ) that divides the universe of discourse "u" of each attribute into "n" intervals according to the number of fuzzy sets after that, the algorithm calculates automatically the width (δi) according to the width (Δ).

3.1.1. Triangular Membership function

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



For "n" linguistic values (fuzzy sets) then the universe of discourse "*U*" which is the attribute divided into "n" intervals with the width between center bi and bi+1 (Δ) where:

$$\Delta = \frac{Xmax - Xmin}{n - 1} \quad \text{where, } 1 \le x \le n+1 \quad \text{with: } Xmin = Xmin - D1, \quad Xmax = Xmax + D2, \quad \text{and } \delta i = (\Delta+1) / 2$$

We can also calculate the parameters a, b and c by the following equations:

Ai = (ai, bi, ci)

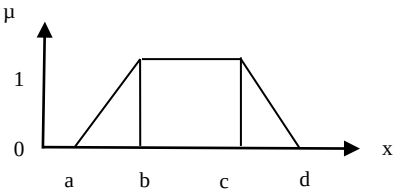
$bi = Xmin + (i - 1) * \Delta$, $ai = bi - \delta i$, $ci = bi + \delta i$

Where:

Δ: the width between center bi and bi+1
Xmax: the maximum value of the attribute
Xmin: the minimum value of the attribute
n: the number of linguistic values (fuzzy sets)
D1: the value subtracted from Xmin to make it integer value
D2: the value added to Xmax to make it integer value
δi: the width between bi and ai, ci

3.1.2. Trapezoidal Membership function

$$\text{Trapezoidal } (X: a, b, c, d) = \begin{cases} 0 & x \le a \\ (x - a) / (b - a) & a \le x \le b \\ 1 & b \le x \le c \\ (d - x) / (d - c) & c \le x \le d \\ 0 & x \ge d \end{cases}$$



For "n" linguistic values (fuzzy sets) then the universe of discourse "u" which is the attribute divided into "n" intervals with the width (Δ) where:

$$\Delta = \frac{Xmax - Xmin}{2\,n - 1} \quad \text{where, } 1 \le x \le n+1 \quad \text{with: } Xmin = Xmin - D1, \quad Xmax = Xmax + D2, \quad \text{and } \delta i = (\Delta+1) / 2$$

We can also calculate the parameters a, b, c and d by the following equations:

Ai = (ai, bi, ci, di)

$bi = Xmin + 2 (i - 1) * \Delta$, $ai = bi - \delta i$, $ci = bi + \Delta$, $di = ci + \delta i$ Where:

Δ: the width between bi and ci, and between ci and bi+1
Xmax: the maximum value of the attribute
Xmin: the minimum value of the attribute
n: the number of linguistic values (fuzzy sets)
D1: the value subtracted from Xmin to make it integer value
D2: the value added to Xmax to make it integer value

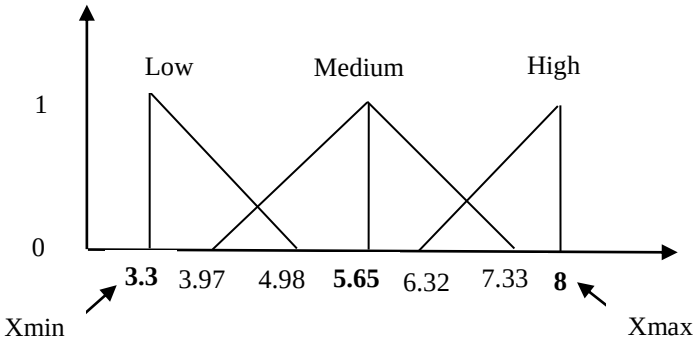
δi: the width between bi and ai, and between ci and di

Example 3.1

Suppose we have the Iris plants dataset (150 records), with information about the attributes *Sepal length (S_L)* in cm, *Sepal width (S_W)* in cm, *Petal length (P_L)* in cm, *Petal width (P_W)* in cm and *Class (RS)* which contain three classes: iris *Setosa*, iris *Versicolour* and iris *Virginica* , where *Sepal length*, *Sepal width*, *Petal length* and *Petal width* are the condition attributes and *Class* is the decision attribute.
Suppose we choose "Triangular Membership function" to linguistic terms and define "3" fuzzy sets: Low, Medium and High. Therefore, to fuzzifing numerical data of attribute *Sepal length (S_L)*, then:

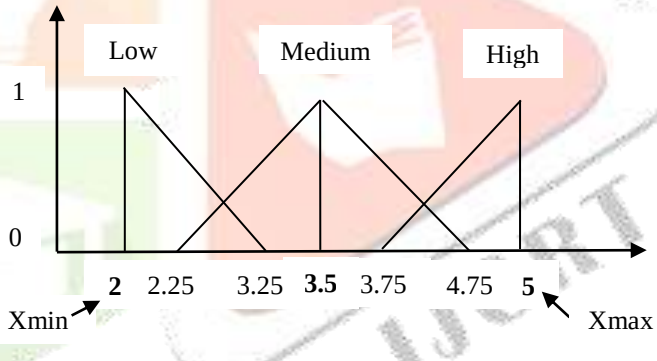
Xmin = 4.3 and Xmax = 7.9 So, D1 = 1 and D2 = 0.1 then:
Xmin = 3.3 and Xmax = 8

The width between bi and bi+1 is:
Δ = 8-3.3/ (3-1) = 2.35
Then, δi = (2.35+1) / 2 = 1.68, therefore:
1- Fuzzy set A1 ⟶ Low, δ1 =1.68
b1 = a1 = 3.3, c1 = 4.98
2- Fuzzy set A2 ⟶ Medium, δ2 =1.68
b2 = 5.65, a2 = 3.97, c2 = 7.33
3- Fuzzy set A3 ⟶ High, δ3 = 1.68
b3 = c3 = 8, a3 = 6.32



Suppose we also choose "Triangular Membership function" to linguistic terms and define "3" fuzzy sets: Low, Medium and High. Therefore, to fuzzifing numerical data of attribute *Sepal width (S_W)*, then:
Xmin = 2.2 and Xmax = 4.4 So, D1 = 0.2 and D2 = 0.6 then:
Xmin = 2 and Xmax = 5

The width between bi and bi+1 is:
Δ = 5-2/ (3-1) = 1.5
Then, δi = (1.5+1) / 2 = 1.25, therefore:
1- Fuzzy set A1 ⟶ Low, δ1 =1.25
b1 = a1 = 2, c1 = 3.25
2- Fuzzy set A2 ⟶ Medium, δ2 =1.25
b2 = 3.5, a2 = 2.25, c2 = 4.75
3- Fuzzy set A3 ⟶ High, δ3 = 1.25
b3 = c3 = 5, a3 = 3.75



Suppose we choose also choose "Triangular Membership function" and define "3" fuzzy sets: Low, Medium and High. Therefore, the algorithm fuzzifing numerical data of two attributes *Petal length (P_L)*, *Petal width (P_W)* in the same manner.
Thus, after fuzzifing numerical data of the all four attributes, the result will be as shown in table 3.

Table 3: Iris data table after fuzzifing the four numerical attributes

Sno	***S_L***	***S_W***	***P_L***	***P_W***	***RS***
1	medium	medium	low	low	setosa
2	medium	medium	low	low	setosa
3	medium	medium	low	low	setosa
4	medium	medium	low	low	setosa
⋮	⋮	⋮	⋮	⋮	⋮
51	high	medium	medium	medium	versicolor
52	medium	medium	medium	medium	versicolor
53	high	medium	medium	medium	versicolor
⋮	⋮	⋮	⋮	⋮	⋮
101	medium	medium	high	high	verginica
102	medium	low	medium	medium	verginica
103	high	medium	high	medium	verginica
⋮	⋮	⋮	⋮	⋮	⋮
150	medium	medium	medium	medium	verginica

Contribution 3.1

Before using our automated Fuzzy-Based Rough Decision Model algorithm on Iris plants dataset, the user should define the parameters of membership functions of these linguistic values from his view which is different from one user to another [11, 21, 31, 32], but by using the proposed automated algorithm the parameters of membership functions of these linguistic values defined automatically that the user determine only the number of fuzzy sets (linguistic values) which is three fuzzy sets in the example then the maximum and minimum values of each attribute are determined automatically after that, the algorithm calculates the width (Δ) that divides the universe of discourse "u" of each attribute into "n" intervals according to the number of fuzzy sets then calculates automatically the width (δi) according to the width (Δ).

3.2. Phase 2: Eliminate Redundant Attributes using Rough Set Theory

There exist some limitations of traditional rough sets theory which restricts its suitability in practice [8, 14, 27, 28], one of which is the inefficiency in computation, which limits its suitability for large data sets in real-world applications. In order to find the reducts, core and dispensable attributes, the rough sets model needs to construct all the equivalent classes based on the attribute values of the condition and decision attributes. This process is very time-consuming, and thus the model is very inefficient and infeasible, and doesn't scale for large data set, which is very common in data mining applications [8, 14]. Further investigation of the problem reveals that most existing rough set models [1, 14, 31, 32] do not integrate with the relational database systems and a lot of computational intensive operations are performed in flat files rather than utilizing the high performance database set operations.
To overcome this problem, a New Rough Sets Model Based on Database Systems has been introduced [6, 14] for this purpose to redefine some concepts of rough set theory such as core attributes and reducts by using relational algebra so that the computation of core attributes and reducts can be performed with very efficient set-oriented database operations, such as the following relational algebra: *Cardinality* (*Card*) to denote the *Count*, and Π for *Projection* operation.

Definition 3.2.1. An attribute $Cj \in C$ is a *core* attribute [6] if it satisfies the following condition:

$Card\ (\Pi\ (C - Cj + D)) \neq Card\ (\Pi\ (C - Cj))$ Where:

Card: the *cardinality* to denote the *count* of attribute, Π: *Projection* operation.

For example, in Table 2, it can be shown that

Card (Π (*C* – *Cj* + *D*)) =*Card* (Π (*Door, Size, Cylinder, Mileage*) = 6, and *Card* (Π (*C* – *Cj*)) = *Card* (Π (*Door, Size, Cylinder*) = 5

Therefore, the attribute *Weight* is a *core* attribute in *C* with respect to attribute *Mileage.*

Definition 3.2.2. An attribute $Cj \in C$ is a *dispensable* attribute [6] in *C* with respect to *D*, if the classification result of each tuple is not affected without using *Cj*, that is,

Card (Π (*C* – *Cj* + *D*)) = Card (Π (*C* – *Cj*))

For example, in Table 2, it can be shown that

Card (Π (*C* – *Cj* + *D*)) = *Card* (Π (*Weight, Size, Cylinder, Mileage*) = 6, and

Card (Π (*C* – *Cj*)) = *Card* (Π (*Weight, Size, Cylinder*) = 6

Thus, *Door* is a dispensable attribute in *C* with respect to attribute *Mileage*.

3.3. Phase 3: Generating Fuzzy Rough Rules and Computation of Accuracy and Confidence values

3.3.1. Generating Fuzzy Rough Rules

After Fuzzifing original information system and determining the reduct attributes we can get decision rules which help decision maker to take the proper decision. We use a new algorithm for extracting fuzzy rough rules from fuzzy table using SQL statements as following:

1- Create temp table

```
SELECT    CURR_REDUCT, D
INTO      TMP_TBL
FROM      T
```

2- Get equivalence classes for current reduct

```
SELECT    CURR_REDUCT
FROM      TMP_TBL
GROUP BY  CURR_REDUCT
```

3- Get decision rule for each equivalence classes

```
SELECT    DISTINCT  D
FROM      TMP_TBL
WHERE     X1 = Y1  AND  X2 = Y2  AND …. Xn = Yn
```

We choose a number of records that the algorithm *randomly* generates the fuzzy rules from them. So, the remaining records will be the test records that calculate the accuracy of those fuzzy rules. After that the algorithm generates a list of Reducts for table 3, we can easily select *Class (RS)* attribute as a decision attribute then the reduct attributes in table 3 will be same four condition attributes *S_L*, *S_W*, *P_L* and *P_W*.

Suppose we choose randomly 75 rows from 150 rows then the algorithm generates 12 fuzzy rules from those records using triangular membership functions as following:

S_L is 'High' And S_W is 'Medium' And P_L is 'High' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'High' And P_W is 'Medium' ==> RS is virginica
S_L is 'High' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'High' ==> RS is virginica
S_L is 'High' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor
S_L is 'Low' And S_W is 'Medium' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'High' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'Low' And P_L is 'High' And P_W is 'Medium' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'High' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor OR virginica
S_L is 'Medium' And S_W is 'Low' And P_L is 'Medium' And P_W is 'Medium' ==> RS is virginica

3.3.2. Confidence (fitness) and support of each rule

In the field of data mining, two measures are often used to evaluate association rules, which are Confidence and Support

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

2- Support (Aq $\rightarrow$ Cq) = $\frac{|D(Aq) \cap D(Cq)|}{|D|} = \frac{\sum \mu Aq(x) \cap \mu Cq(x)}{\text{No. of all tuples}}$ 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

For the pervious fuzzy rough rules that generated from table 3, the algorithm will automatically calculate the confidence and the accuracy of each fuzzy rough rule which the confidence calculated using the Average operator that takes the average confidence value of the fuzzy sets of each rule then it calculates the total accuracy of all rules as the following results in the table 4.

Table 4: *Confidence* and *Accuracy* of each Fuzzy rough Rule before running Genetic algorithm

Fuzzy rules	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	Total	Accuracy %
frequency	1	9	1	2	3	2	1	5	28	1	15	7	75	
Confidence	0.82	0.6	0.76	0.66	0.84	0.83	0.63	0.71	0.7	0.66	0.55	0.58		
Accuracy	2	0	0	0	1	0	1	1	15	2	19	1	42	0.56

3.4. Phase 4: Using genetic algorithm for the Fuzzy rough rules

The last phase is running the Genetic algorithm on the pervious fuzzy rough rules. We represent the three Fuzzy sets in each rule as a one chromosome with a number from 0 to 9. Suppose we represent each fuzzy set of the pervious fuzzy rough rules by integer number forming a gene on chromosome (individual) of fixed length that forms a population of fixed number of chromosomes as following:

1: Low, 2: Medium, 3: High, 0: Do not care

After that, we set the accuracy value condition that we need to reach. Suppose we set the accuracy value to be 95%
Then the algorithm generates another new 12 fuzzy rules using triangular membership functions as following:

S_L is 'High' And S_W is 'Medium' And P_L is 'High' And P_W is 'High' ==> RS is virginica
S_L is 'High' And S_W is 'Medium' And P_L is 'High' And P_W is 'Medium' ==> RS is virginica
S_L is 'High' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'High' ==> RS is virginica
S_L is ' High' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor
S_L is 'Low' And S_W is 'Medium' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'High' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'Low' And P_L is 'High' And P_W is 'Medium' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'High' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Low' And P_W is 'Low' ==> RS is setosa
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'High' ==> RS is virginica
S_L is 'Medium' And S_W is 'Medium' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor OR virginica
S_L is 'Medium' And S_W is 'Low' And P_L is 'Medium' And P_W is 'Medium' ==> RS is versicolor OR virginica

- After that, the accuracy of those new 12 rules from randomly 75 rows from 150 rows will automatically calculated to be 0.95 instead of 0.56
- Suppose we choose randomly 100 rows from 150 rows, then the accuracy of those new 13 rules will automatically calculate to be 0.96 instead of 0.68
- Suppose we choose randomly 130 rows from 150 rows, then the accuracy of those new 14 rules will automatically calculate to be 0.95 instead of 0.7

The results of the three cases of Iris Dataset after running Genetic will be as following in the table 5:

Table 5: *Confidence* and *Accuracy* after running Genetic algorithm of three cases of Iris Data

	Fuzzy rules	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	Total	Accuracy %
Case 1 75 rows	frequency	1	***7***	1	2	3	2	1	5	28	1	15	***9***			***75***	
	Confidence	0.82	***0.71***	0.76	0.66	0.84	0.83	0.63	0.71	0.7	0.66	0.55	***0.58***				
	Accuracy	2	***6***	0	0	1	0	1	1	15	2	19	***24***			**71**	**0.95**
Case 2 100 rows	frequency	1	5	1	2	3	2	1	***19***	5	4	28	1	***28***		***100***	
	Confidence	0.82	0.7	0.8	0.7	0.84	0.8	0.63	***0.58***	0.71	0.6	0.7	0.66	***0.55***			
	Accuracy	2	5	0	0	1	0	1	***10***	1	5	15	2	***6***		**48**	**0.96**
Case 3 130 rows	frequency	1	3	8	1	2	2	2	2	***23***	6	8	41	3	***28***	***130***	
	Confidence	0.85	0.82	0.71	0.76	0.66	0.84	0.83	0.63	***0.58***	0.71	0.6	0.7	0.66	***0.55***		
	Accuracy	0	0	2	0	0	2	0	0	***6***	0	1	2	0	***6***	**19**	**0.95**

Another example is Data_User_Modeling_Dataset (145 records) from UCI Machine learning, which has 5 condition attributes STG, SCG, STR, LPR, PEG and one decision attributes UNS.

- Suppose we choose randomly 50 rows from 145 rows, then the accuracy of those new 27 rules will automatically calculate to be 0.46 instead of 0.18
- Suppose we choose randomly 75 rows from 145 rows, then the accuracy of those new 41 rules will automatically calculate to be 0.57 instead of 0.21
- Suppose we choose randomly 100 rows from 145 rows, then the accuracy of those new 52 rules will automatically calculate to be 0.62 instead of 0.29
- Suppose we choose randomly 125 rows from 145 rows, then the accuracy of those new 64 rules will automatically calculate to be 0.75 instead of 0.33

The results of the four cases of Data_User_Modeling_Dataset after running Genetic will be as following in the table 6:

Table 6: *Confidence* and *Accuracy* after running Genetic algorithm of four cases of DUM Data

	Fuzzy rules	R1	R2	R3	….	R27	R28	….	R41	R42	….	R52	R53	….	R64	Total	Accuracy %
Case 1 50 rows	frequency	1	***1***	1	….	5										***50***	
	Confidence	0.66	***0.46***	0.59	….	0.38											
	Accuracy	0	***5***	1	….	2										**44**	**0.46**
Case 2 75 rows	frequency	1	1	1	….	1	3	….	2							***75***	
	Confidence	0.69	0.61	0.66	….	0.48	0.43	….	0.43								
	Accuracy	0	0	0	….	0	3	….	3							**43**	**0.57**
Case 3 100 rows	frequency	1	1	1	….	1	1	….	4	***1***	….	4				***100***	
	Confidence	0.7	0.6	0.7	….	0.43	0.48	….	0.46	***0.46***	….	0.42					
	Accuracy	0	0	0	….	0	0	….	2	***5***	….	2				**28**	**0.62**
Case 4 125 rows	frequency	1	1	1	….	5	3	….	2	2	….	1	***1***	….	4	***125***	
	Confidence	0.69	0.65	0.7	….	0.41	0.51	….	0.51	0.47	….	0.56	***0.48***	….	0.42		
	Accuracy	0	0	0	….	0	0	….	0	0	….	1	***4***	….	2	**15**	**0.75**

3.5 The performance curve

The performance curve of values for maximum **Accuracy** for each generation of the running program can be represented by Figure 3.5 and 3.6

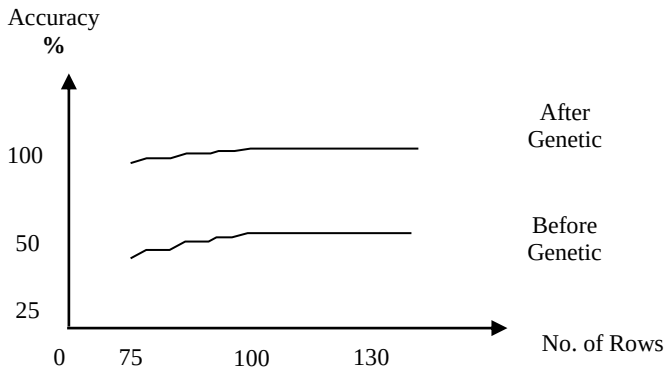


Figure 3.5: The performance curve of the Iris plants Dataset

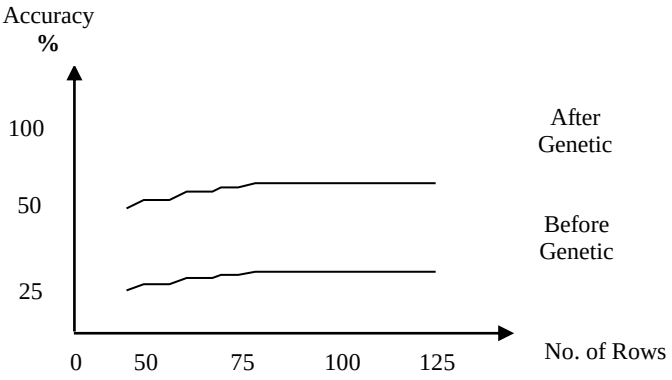


Figure 3.6: The performance curve of the Data_User_Modeling_Dataset

Contribution 3.2

Our Genetic-Based Fuzzy Rough Decision Model algorithm is more efficient than the Traditional Genetic fuzzy and rough models [33, 35, 36, 37], because our algorithm calculates automatically the confidence (fitness) of each rule by using the Average operator that takes the

average confidence value of the fuzzy sets of each rule. Then the algorithm also calculates automatically the accuracy of each fuzzy rule. Then, the rules that have a high accuracy are called *Strong rules*. After that, the algorithm using Genetic algorithm technique to generates another new fuzzy rules then it automatically calculates the accuracy of those rules which will be higher than the old rules before using Genetic algorithm.

4. Conclusion

Rough sets theory has been applied successfully in many disciplines. One of the major limitations of the traditional rough sets model is that it assumes that all attribute values are discrete. A real world data always contains mixed types of data such as continuous valued, symbolic data, etc. therefore all numerical or continuous data should converted to discretized data, here "Fuzzy Logic" can solve this problem to reduce information overload in a Fuzzy-Based Rough Decision Model.

One drawback of traditional Fuzzy-Based Rough Decision Model is that the linguistic values (fuzzy sets) for numerical values of each attribute has to be determining by the membership functions of these linguistic terms which, the user should define the parameters of membership functions of these linguistic values from his view which is different from one user to another. Therefore, we propose a new automated Fuzzy-Based Rough Decision Model algorithm that can define the parameters of membership functions of these linguistic values automatically by determine the number of fuzzy sets and the maximum and minimum values of each attribute then the algorithm finds the width (Δ) that divides the universe of discourse "*U*" of each attribute into "n" intervals according to the number of fuzzy sets then the algorithm calculates automatically the width (δi) according to the width (Δ).

Another drawback of the traditional rough sets model in the real applications is the inefficiency in the computation of core attributes and the generation of reducts. Most existing rough set models, do not integrate with database systems and a lot of computational intensive operations such as discernibility relations computation, core attributes search, reduct generation, and rule induction are performed on flat files, which limits their applicability for large data sets in data mining applications.

In order to improve the efficiency of computing core attributes and reducts, a New Rough Sets Model Based on Database Systems has been introduced for this purpose [Hu, X., Lin, T., 2004], which redefine the core attributes and reducts based on relational algebra such as *Cardinality*, *Projection*, and *Selection* and so on to take advantages of the very efficient set-oriented database operations.

Our algorithm automatically calculates the confidence (fitness) and the accuracy of each fuzzy rough rule then it calculates the total accuracy value of all linguistic rules.

After that, the algorithm using Genetic algorithm technique to generates another new fuzzy rules then it automatically calculates the accuracy of those rules which will be higher than the old rules before using Genetic algorithm.

5. Future work

Depending on the choice of parameters, association fuzzy rule algorithms can generate an extremely large number of rules which lead algorithms to suffer from long execution time and huge memory consumption, So we will propose a Top k Rules algorithm to discover the *top-k* rules before using Genetic algorithm which having the highest support and confidence, where *k* is set by the user. This algorithm discovers all rules that have a support and confidence respectively higher or equal to user-defined thresholds minimum support *minsup* and minimum confidence *minconf*.

References

[1] Bazan, J., Nguyen, H., Nguyen, S., Synak, P., Wroblewski, J., Rough set algorithms in classification problems, *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*, L. Polkowski, T. Y. Lin, and S. Tsumoto (eds), 49-88, Physica-Verlag, Heidelberg, Germany, 2000

[2] Cercone N., Ziarko W., Hu X., Rule Discovery from Databases: A Decision Matrix Approach, *Proc. Of ISMIS*, Zakopane, Poland, 653-662, 1996

[3] Q. Shen, J. Richard**,** Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, Pattern Recognition 37(2004) 1351–1363.

[4] Fernandez-Baizan, A., Ruiz, E., Sanchez, J., Integrating RDMS and Data Mining Capabilities Using Rough Sets, *Proc. IMPU*, Granada, Spain, 1996

[5] Bhatt R B, Gopal M. On fuzzy-rough sets approach to feature selection. Pattern Recognition Letters, 2005, 26(7): 965–975.

[6] Hu, X., Lin, T., Han, J, " A New Rough Sets Model Based on Database Systems" , Fundamenta Information 59 no. 2-3 pp.135-152 ., 2004

[7] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. International Journal of General Systems, 1990, 17(2–3): 191.

[8] Hu, X., Cercone N., Han, J., Ziarko, W, GRS: A Generalized Rough Sets Model, in Data Mining, *Data Mining, Rough Sets and Granular Computing*, T.Y. Lin, Y.Y.Yao and L. Zadeh (eds), Physica-Verlag, 447-460, 2002

[9] Drwal G., Sikora M.: Fuzzy Decision Support System with Rough Set Based Rules Generation Method. In:Rough Sets and Current Trends in Computing, LNAI 3006. Springer-Verlag (2004) 727 733.

[10] John,G., Kohavi, R., Pfleger,K., Irrelevant Features and the Subset Selection Problem, *Proc. ICML*, 121-129, 1994

[11] Qiang He a, Congxin Wua, Degang Chen b, Suyun Zhao.: Fuzzy rough set based attribute reduction for information systems with fuzzy decisions, *Knowledge-based systems* 24 (2011) 689-696.

[12] Kira,K., Rendell,L.A. The feature Selection Problem: Traditional Methods and a New Algorithm, *Proc. AAAI*, MIT Press, 129-134, 1992

[13] Zhi, W.W, J.S. Mi and W.X. Zhang "Generalized fuzzy Rough Sets", Information Sciences, vol.151, p.p. 263–282, 2003.

[14] Kumar A., New Techniques for Data Reduction in Database Systems for Knowledge Discovery Applications, *Journal of Intelligent Information Systems*, 10(1), 31-48, 1998

[15] Wang, X.Z., E.C.C. Tsangb, S. Zhao, D. Chen and D.S. Yeung, "Learning fuzzy rules from fuzzy samples based on rough set technique", Information Sciences, vol. 177, p.p. 4493–4514, 2007.

[16] Lin T.Y., Cercone, N. (eds), Rough Sets and Data Mining: Analysis of Imprecise Data, Kluwer Academic Publisher, 1997
[17] A.M. Radzikowska, E.E. Kerre, A comparative study of rough sets, Fuzzy Sets and Systems 126 (2002) 137–155.
[18] Lin T.Y., Yao Y.Y. Zadeh L. (eds), Data Mining, Rough Sets and Granular Computing, Physica-Verlag, 2002
[19] Q. Shen, A. Chouchoulas, A rough fuzzy approach for generating classification rules, Pattern Recognition 35 (2002) 2425–2438.
[20] Liu, H., Motoda., H. (eds), Feature Extraction Construction and Selection: A Data Mining Perspective. Kluwer Academic Publisher, 1998
[21] Jiuping Xu, Lihui Zhao.: A multi-objective decision-making with fuzzy rough coefficients to the inventory problem. *Information sciences*,180(2010) 679-696
[22] Modrzejewski, M., Feature Selection Using Rough Sets Theory, *Proc. ECML*, 213-226, 1993
[23] Nguyen, H., Nguyen, S., Some efficient algorithms for rough set methods, *Proc. IPMU* Granada, Spain, 1451-1456, 1996
[24] Lirong J, Liu S. Extension of rough sets model for probabilistic decision analysis from rough fuzzy decision tables. Journal of Southeast University, 2006b, 2(5): 246–250**.**
[25] Pawlak Z., Rough Sets, *International Journal of Information and Computer Science*, 11(5), 341-356, 1982
[26] Pawlak Z., Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1992
[27] Polkowski, L., Skowron, A., Rough mereology, *Proc. ISMIS*, Charlotte, NC, 85-94, 1994
[28] Polkowski, L., Skowron, A., Rough mereology: A new paradigm for approximate reasoning, *J. of Approximate Reasoning*, 15(4), 333-365, 1996
[29] Skowron, A., Rauszer C., The Discernibility Matrices and Functions in Information Systems, *Intelligent Decision Support - Handbook of Applications and Advances of the Rough Sets Theory*, K. Slowinski (ed), Kluwer, Dordrecht, 331-362, 1992
[30] Yifan Wang, Mining stock price using fuzzy rough set system, Expert Systems with Applications 24 (2003) 13–23.
[31] Tzung-Pei Hong a, Li-Huei Tseng b, Been-Chian Chien c.: Mining from Incomplete quantitative data by fuzzy rough sets, Expert System with Application 37 (2009) 2644-2653.
[32] Yan-Qing Yao, Ju-Sheng Mi, Zhou-Jun Li. Attribute reduction based on Generalized fuzzy evidence theory in fuzzy sets and systems. Fuzzy Sets and Systems, (2011) 509-517.
[33] Cordon, O., Herrera, F., Hoffmann, F., Magdalena, L. (2001): Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases. World Scientific Publishers, Singapore.
[34] I.H. Toroslu, Y. Arslanoglu, Genetic algorithm for the personnel assignment problem with multiple objectives, Information Sciences 177 (2007) 787–803.
[35] Wen-Yau Liang, Chun-Che Huang.: The generic genetic algorithm incorporates with rough set theory – An application of the web services composition. Expert Systems with Applications 36 (2009) 5549-5556.
[36] Marek Sikora.: Fuzzy Rules Generation Method for Classification Problems Using Rough Sets and Genetic Algorithms. Computer Science, (2005), 44-100.
[37] Mohammad Lutfi Othman, Ishak Aris, Mohammad Ridzal Othman, Harussaleh Osman. Rough-Set-and-Genetic-Algorithm based data mining and rule Quality Measure to hypothesize distance protective relay operation characteristics from relay even report.ElectricalpowerandEnergySystems33 (2011)1437-1