Chapter 3

Rough Set and Fuzzy Rough Set Based Approaches

3.1 Introduction

The purpose of this chapter is to familiarize the reader with the terms and the background related to the Rough Set and Fuzzy Rough Set given in the literature.

In the previous chapter, it has been discussed that keeping classification accuracy as fitness function makes feature selection task computationally intensive and it does not allow to take benefit of interdependency present among the objects.

In view of the above, alternatives to the classification accuracy based fitness function, should be investigated. Alternatives to the fitness function are information gain [63], correlation [64], mRMR [65], Fisher score [66,67] and distance based measure [68,69]. Literature [2–4,9,30,36] suggests that Rough Set and Fuzzy Rough Set based fitness functions are found better in capturing the interdependencies of the features of any dataset.

3.2 Rough Set Theory (RST) Based Feature Selection [1–4]

In this section rough set based feature selection has been discussed [1]. This technique does not need any other information or parameter beside the data provided. An example dataset is given in Table 3.2 to illustrate the procedure of rough set based feature selection

Table 3.1: Sample dataset 1

Object	a	b	c	q
1	-0.4	-0.3	-0.5	0
2	-0.4	0.2	-0.1	1
3	-0.3	-0.4	-0.3	0
4	0.3	-0.3	0	1
5	0.2	-0.3	0	1
6	0.2	0	0	0

Table 3.2: Sample Dataset 2: discretized version of sample dataset 1

Object	a	b	c	q
1	E	F	M	H
2	E	J	G	I
3	F	E	F	H
4	K	F	H	I
5	J	F	H	I
6	J	H	H	H

[2,3].

This fact can be explained with the help of an example dataset given in Table 3.1. In this table there are six objects given as rows. First three columns (columns a,b and c) represent features for each object, last column (column q) represents decision variable, decision feature or class label.

Table 3.2 is the analogous and discretized version of Table 3.1. Table 3.2 will be used to illustrate an example of rough set measure as follows.

Let $I = (U, A)$ be an information system, where U is a non-empty set of finite objects and A is the nonempty finite set of features including the class labels, such that $a : U \rightarrow V_a$ for every $a \in A$. V_a is the set of values that feature a may take. P is a set of conditional feature and Q is a set of class label feature set. With any $P \subseteq A$, there is an associated equivalence relation $IND(P)$ [1,2,4].

$$IND(P) = (x, y) \in U^2 | \forall a \in P, a(x) = a(y) \tag{3.1}$$

The partition of U, which is generated by $IND(P)$, is denoted $U/IND(P)$ (or U/P) and can be calculated as follows.

$$U/IND(P) = \otimes U/IND(a) | a \in P \tag{3.2}$$

Where $\otimes$ is specifically defined as follows for sets A and B.

$$A \otimes B = X \cap Y | X \in A, Y \in B, X \cap Y \neq \phi \tag{3.3}$$

If (x,y) belongs $IND(P)$, then x and y are indiscernible by features from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_p$. For example, using Table 3.2, if $P = \{a, c\}$, the objects 5, 6 are indiscernible. $IND(P)$ creates the following partition of U.

$$U/IND(P) = U/IND(a) \otimes U/IND(c)$$

$$= \{\{1, 2\}, \{3\}, \{4\}, \{5, 6\}\} \otimes \{\{1\}, \{2\}, \{3\}, \{4, 5, 6\}\}$$
$$= \{\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6\}\}$$

Let $X \subseteq U$ and X can be approximated using only the information contained within P by constructing the P-lower and P-upper approximations of X.

$$\underline{P}X = \{x \in U | [x]_P \subseteq X\} \tag{3.4}$$

$$\overline{P}X = \{x \in U | [x]_P \cap X \neq \phi\} \tag{3.5}$$

The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called a Rough Set. Let P and Q be the set of features inducing equivalence relations over U, then the positive, negative, and boundary regions can be defined as follows.

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{P}X \tag{3.6}$$

$$NEG_P(Q) = U - \bigcup_{X \in U/Q} \overline{P}X \tag{3.7}$$

$$BND_P(Q) = \bigcup_{X \in U/Q} \overline{P}X - \bigcup_{X \in U/Q} \underline{P}X \tag{3.8}$$

For example it is obvious that the lower approximation of $P = \{a, c\}$ against the class {q} is {1,3} and {2,4} for class labels H and I respectively.

$$POS_{\{a,c\}}(\{q\}) = \bigcup\{\{1,3\},\{2,4\}\}$$

$$POS_{\{a,c\}}(\{q\}) = \{1,2,3,4\}$$

This means that objects 5 and 6 cannot be classified because the information that would make them discernible is not there. In other words we find that considering only features a and c of object 5 and object 6 the corresponding class labels are different.

3.2.1 Dependency Measure

An important issue in data analysis is discovering dependencies between the features of the dataset. If all the feature values of a set from Q are uniquely determined by the feature values of a set of features P, then the dependency measure will be 1 (maximum).

In rough set theory, dependency is defined in the following way.
For $P, Q \subset A$, k ($0 \leq k \leq 1$) is degree of dependency of Q on P (denoted $P \Rightarrow_k Q$), if

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|}. \tag{3.9}$$

As we know, Q is class label feature and P is set of conditional features. If k = 1 then Q totally depends on P, else if $0 < k < 1$, Q partially depends on P. If k = 0 then Q does not depends on P.

For the dataset under consideration (Table 3.2), the degree of dependency of class label feature $\{q\}$ on the attributes $\{a, c\}$ is

$$\gamma_{\{a,c\}}(\{q\}) = \frac{|POS_{\{a,c\}}(\{q\})|}{|U|}$$

$$= |1,2,3,4|/|1,2,3,4,5,6|$$

$$= 4/6$$

Table 3.3: Rough Dependency Measure

P	Rough Dependency Measure
{a}	$\gamma_{\{a\}}(Q) = 0.33$
{b}	$\gamma_{\{b\}}(Q) = 0.5$
{c}	$\gamma_{\{c\}}(Q) = 0.5$
{a,b}	$\gamma_{\{a,b\}}(Q) = 1$
{b,c}	$\gamma_{\{b,c\}}(Q) = 1$
{a,c}	$\gamma_{\{a,c\}}(Q) = 0.66$

$$= 0.66$$

Similarly dependency degree for other features and their combinations have been computed and shown in the Table 3.3 using the same procedure.

The reduction of features is achieved by comparing equivalence relations generated by sets of features. Features are removed in such a way that the reduced set provides the same predictive capability for the decision feature i.e. class label feature as it is for the original feature set.

A reduct R is defined as a minimal subset of the initial feature set C such that for a given set of feature D (class labels), $\gamma_R(D) = \gamma_C(D)$. A given dataset may have many reduct sets. The intersection of all the reduct sets are called *core*.

3.3 Fuzzy Rough Set Based Feature Selection [5–10]

3.3.1 Fuzzy Rough Set Feature Selection

The combination of rough sets and fuzzy sets as proposed by [5] give rise to the notion of fuzzy rough sets. This provides an effective means of overcoming the problem of the discretization and can be directly applied to the reduction of numerical or continuous attributes [10]. The rough set feature selection process described previously can only operate effectively with dataset containing discrete values. Further, they are unable to handle noisy data. As most datasets contain real-valued features, it is necessary to perform discretization of given data if rough set based feature selection is to be applied. To handle real valued features, fuzzy rough set feature selection methods have been suggested [8] as described in the following paragraph.

Fuzzy Equivalence classes

As crisp equivalence classes are central to rough set approach, fuzzy equivalence classes are central to fuzzy rough set approach [6,8]. The concept of crisp equivalence classes can be extended by introducing fuzzy similarity relation, S, on the universe, which determines the extent to which two elements are similar in S [8]. Thus fuzzy similarity relation can capture uncertainty and vagueness present in the real valued features. The usual properties of reflexivity ($\mu_S(x,x) = 1$), symmetry ($\mu_S(x,y) = \mu_S(y,x)$), and T-transitivity ($\mu_S(x,z) \geq \mu_S(x,y) \wedge_T \mu_S(y,z)$) hold.

The family of normal fuzzy sets produced by a fuzzy partioning of the Universe of Discourse (UoD) can play the role of fuzzy equivalence classes [70]. The crisp partitioning of Universe of Discourse, U, by the features in Q, U/Q, can be thought of as degenerated fuzzy sets, where elements belonging to the class have a membership of one, otherwise zero. The crisp equivalence classes can be extended to fuzzy equivalence classes. If we allow objects to assume membership values between 0 and 1, U/Q is applicable to fuzzy partitions also. For the research presented here, Fuzzy sets (corresponding to fuzzy equivalence classes) are derived using simple fuzzification preprocessor. The preprocessor used the statistical properties of data.

Fuzzy Rough Sets

In the crisp case, elements that belong to lower approximation (i.e. have a membership value 1) are said to belong to the approximated set with absolute certainty.

In the fuzzy rough case elements may have membership in range [0,1], which allows greater flexibility in handling uncertainty and vagueness. Fuzzy-rough sets incorporate the distinct concepts of vagueness (for fuzzy sets) and indiscernibility (for rough sets), both occuring as a result of uncertainty in knowledge [8, 9].

The tuple $< \mu_{\underline{P}X}, \mu_{\overline{P}X} >$ is called a fuzzy rough set. Fuzzy P-lower approximation, $\mu_{\underline{P}X}$ and fuzzy P-upper approximation, $\mu_{\overline{P}X}$ can be computed using the definition of fuzzy rough sets with the following formula suggested in [6].

$$\mu_{\underline{P}X}\left(F_i\right) = \inf_x max\left\{1 - \mu_{F_i}\left(x\right), \mu_X\left(x\right)\right\} \forall i, and \tag{3.10}$$

$$\mu_{\overline{P}X}\left(F_i\right) = \sup_x min\left\{\mu_{F_i}\left(x\right), \mu_X\left(x\right)\right\} \forall i, \tag{3.11}$$

where F_i is a fuzzy equivalence class, and X is the (fuzzy) concept to be approximated.

3.3.2 Fuzzy Rough Dependency Measure

The crisp positive region in traditional rough set theory is defined as the union of the lower approximations. By the extension principle [71], the membership of an object $x \in U$ belonging to the fuzzy positive region can be defined by.

$$\mu_{POS_P(Q)}\left(x\right) = \sup_{X \in U/Q} \mu_{\underline{P}X}\left(x\right). \tag{3.12}$$

Further the fuzzy rough dependency measure can be defined as follows.

$$\gamma_P^{'}(Q) = \frac{\left|\mu_{POS_P(Q)}\left(x\right)\right|}{|U|} = \frac{\Sigma_{x \in U}\mu_{POS_P(Q)}\left(x\right)}{|U|} \tag{3.13}$$

This fuzzy rough dependency measure can be used as a fitness function for feature selection. The set of features giving maximum value of fuzzy rough dependency measure, is finalized as optimal reduct.

Difficulties with Fuzzy Rough Feature Selection

Fuzzy Rough Feature Selection method has several inherent problem [8].

(i) Complexity of calculating the Cartesian product of fuzzy equivalence classes becomes high for large feature subsets.

(ii) If the number of fuzzy sets per feature is n, $n^{|R|}$ equivalence classes must be considered per feature for any candidate reduct, R. To reduce the computational effort required to calculate fuzzy lower approximations for large dataset, a compact computational domain based on some of the properties of fuzzy connectives are reported in [7].

(iii) Further, another problem is that sometimes the fuzzy lower approximation might not be a subset of the fuzzy upper approximation, which is undesirable and meaningless.

(iv) Cartesian product of fuzzy equivalence classes might not result in a family of fuzzy equivalence classes.

3.4 Lower approximation based Fuzzy Rough Feature Selection (L-FRFS) [8, 11]

Due to the problem mentioned in the previous section a new feature selection technique, named as Fuzzy lower approximation based Fuzzy Rough Feature Selection (L-FRFS) has been introduced in literature [8, 11].

L-FRFS uses a fuzzy partitioning of the input space in order to determine fuzzy equivalence classes. Alternative definition for fuzzy lower and fuzzy upper approximation are as under.

$$\mu_{\underline{R_P}X}\left(x\right) = \inf_{y\in U} I\left(\mu_{R_P}\left(x,y\right), \mu_X\left(y\right)\right) \tag{3.14}$$

$$\mu_{\overline{R_P}X}\left(x\right) = \sup_{y\in U} I\left(\mu_{R_P}\left(x,y\right), \mu_X\left(y\right)\right) \tag{3.15}$$

.

Here, I is a fuzzy implicator $(min(1-x+y,1))$ and R_p is the fuzzy similarity relation induced by the subset of features P as described below [8, 9].

$$\mu_{R_P}\left(x,y\right) = T_{a\in P}\left\{\mu_{R_a}\left(x,y\right)\right\} \tag{3.16}$$

where, T is a t-norm $(max(x+y-1,0))$, $\mu_{R_a}\left(x,y\right)$ is the degree to which objects x and y are similar for feature a.

Many fuzzy similarity relations can be constructed as below.

$$\mu_{R_a}\left(x,y\right) = 1 - \frac{|a\left(x\right) - a\left(y\right)|}{|a_{max} - a_{min}|}, \tag{3.17}$$

$$\mu_{R_a}\left(x,y\right) = exp\left(-\frac{\left(a\left(x\right) - a\left(y\right)\right)^2}{2\sigma_a^2}\right), \tag{3.18}$$

$$\mu_{R_a}(x,y) = max(min(\frac{a(y) - a(x) + \sigma_a}{\sigma_a}, \frac{a(x) - a(y) + \sigma_a}{\sigma_a}), 0), \tag{3.19}$$

where σ is the standard deviation of attribute a. Fuzzy transitive closure must be computed for each feature because the above similarity relations do not necessarily show T-transitivity. The combination of features also need to retain T-transitivity.

Fuzzy positive region may be defined as,

$$\mu_{POS_{R_P}(Q)}(x) = \sup_{X \in U/Q} \mu_{\underline{R_P}X}(x) \tag{3.20}$$

.

The fuzzy rough dependency measure, $\gamma'_P(Q)$, may be defined as follows.

$$\gamma'_P(Q) = \frac{\Sigma_{x \in U} \mu_{POS_{R_P}(Q)}(x)}{|U|} \tag{3.21}$$

A fuzzy rough based reduct, R, can be defined as a subset of features that preserves the degree of dependency of the entire dataset, i.e. $\gamma'_R(D) = \gamma'_C(D)$ where C is the set of conditional features.

Further, these computations can be done on real valued features. Let us take an example dataset shown in Table 3.1 for sample calculation of L-FRFS based dependency measure.

The fuzzy connectives chosen throughout this thesis and in this example are Lukasiewicz fuzzy implicator ($min(1 - x + y, 1)$) and Lukasiewicz t-norm ($max(x + y - 1, 0)$). Using the method discussed, its fuzzy rough dependency measure has been calculated and tabulated in Table 3.4 along with corresponding rough dependency measure for corresponding discretized data of Table 3.2. To understand the process sample calculation using above method has been given in subsequent subsection.

Table 3.4: Rough and Fuzzy Rough Dependency Measure

P	Dependency Measure based on	
	Rough Set	Fuzzy Rough Set
{a}	$\gamma_{\{a\}}(Q) = 0.33$	$\gamma'_{\{a\}}(Q) = 0.1002$
{b}	$\gamma_{\{b\}}(Q) = 0.5$	$\gamma'_{\{b\}}(Q) = 0.3597$
{c}	$\gamma_{\{c\}}(Q) = 0.5$	$\gamma'_{\{c\}}(Q) = 0.4078$
{a,b}	$\gamma_{\{a,b\}}(Q) = 1$	$\gamma'_{\{a,b\}}(Q) = 1$
{b,c}	$\gamma_{\{b,c\}}(Q) = 1$	$\gamma'_{\{b,c\}}(Q) = 1$
{a,c}	$\gamma_{\{a,c\}}(Q) = 0.66$	$\gamma'_{\{a,c\}}(Q) = 0.5501$

Similarly dependency degree for other features and their combinations can be computed as shown in Table 3.4 using the same procedure. In Table 3.4, comparing the values of dependency measures using rough set method (γ) and fuzzy rough set method (γ'), for features sets, it is noted that dependency measure in case of fuzzy rough set is lower than the corresponding values in case of rough sets.

It is because of the dependency measure (γ), in case of rough set can have a resolution of 1/(number of objects) hence it can not capture dependency below this resolution whereas the resolution of fuzzy dependency measure ($\gamma^{'}$) is dependent actually on the feature values and is, thus, able to capture higher resolution in the dependency measure of the feature sets.

3.4.1 Computation of Fuzzy Rough Dependency Measure: Sample Calculation

Using the fuzzy similarity relation defined in Equation (3.19), the resulting few relations are as follows.

$$\mu_{R_a}(x,y) = \begin{bmatrix} 1 & 1 & 0.6994 & 0 & 0 & 0 \\ 1 & 1 & 0.6944 & 0 & 0 & 0 \\ 0.6994 & 0.6994 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.6994 & 0.6994 \\ 0 & 0 & 0 & 0.6994 & 1 & 1 \\ 0 & 0 & 0 & 0.6994 & 1 & 1 \end{bmatrix}$$

$$\mu_{R_b}(x,y) = \begin{bmatrix} 1 & 0 & 0.5683 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.1367 \\ 0.5683 & 0 & 1 & 0.5683 & 0.5683 & 0 \\ 1 & 0 & 0.5683 & 1 & 1 & 0 \\ 1 & 0 & 0.5683 & 1 & 1 & 0 \\ 0 & 0.1367 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mu_{R_c}(x,y) = \begin{bmatrix} 1 & 0 & 0.0355 & 0 & 0 & 0 \\ 0 & 1 & 0.0355 & 0.5178 & 0.5178 & 0.5178 \\ 0.0355 & 0.0355 & 1 & 0 & 0 & 0 \\ 0 & 0.5178 & 0 & 1 & 1 & 1 \\ 0 & 0.5178 & 0 & 1 & 1 & 1 \\ 0 & 0.5178 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Equation (3.16) can be used for finding other fuzzy similarity relation induced by the subset of features P. For example if P ={a,b}

$$\mu_{R_{ab}}\left(x,y\right) = T\left\{\mu_{R_a}\left(x,y\right), \mu_{R_b}\left(x,y\right)\right\}$$

In order to compute the dependency measure, the very first step is computing the fuzzy lower approximation of each class for each feature. Considering feature b and the decision class {1,3,6} in the example dataset 2.

$$\mu_{\underline{R_b}\{1,3,6\}}\left(x\right) = \inf_{y \in U} I\left(\mu_{R_b}\left(x,y\right), \mu_{\{1,3,6\}}\left(y\right)\right)$$

For object 1, using above formula lower approximation is;

$$\mu_{\underline{R_b}\{1,3,6\}}\left(1\right) = \inf_{y \in U} I\left(\mu_{R_b}\left(1,y\right), \mu_{\{1,3,6\}}\left(y\right)\right)$$

$$= inf\{I(1,1), I(0,0), I(0.5683,1), I(1,0), I(1,0), I(0,1)\}$$

$$= inf\{1,1,1,0,0,1\}$$

$$= 0$$

The fuzzy lower approximation for the remaining objects considering decision class {1,3,6} can be calculated in the similar way. Finally, the fuzzy lower approximations considering decision class {1,3,6} are as follows.

$$\mu_{\underline{R_b}\{1,3,6\}}\left(1\right) = 0$$

$$\mu_{\underline{R_b}\{1,3,6\}}\left(2\right) = 0$$

$$\mu_{\underline{R_b}\{1,3,6\}}\left(3\right) = 0.4317$$

$$\mu_{\underline{R_b}\{1,3,6\}}\left(4\right) = 0$$

$$\mu_{\underline{R_b}\{1,3,6\}}\left(5\right) = 0$$

$$\mu_{\underline{R_b}\{1,3,6\}}\left(6\right) = 0.8633$$

Similarly for decision class {2,4,5}, the fuzzy lower approximation for each of the object would be.

$$\mu_{\underline{R_b}\{2,4,5\}}(1) = 0$$

$$\mu_{\underline{R_b}\{2,4,5\}}(2) = 0.8633$$

$$\mu_{\underline{R_b}\{2,4,5\}}(3) = 0$$

$$\mu_{\underline{R_b}\{2,4,5\}}(4) = 0$$

$$\mu_{\underline{R_b}\{2,4,5\}}(5) = 0$$

$$\mu_{\underline{R_b}\{2,4,5\}}(6) = 0$$

Using Equation (3.20), fuzzy positive region for the object 1, can be computed as follows.

$$\mu_{POS_{R_b}(Q)}(1) = max(\mu_{\underline{R_b}\{1,3,6\}}(1), \mu_{\underline{R_b}\{2,4,5\}}(1))$$

$$= max(0, 0)$$

$$= 0$$

Thus the positive regions for each of the objects are

$$\mu_{POS_{R_b}(Q)}(1) = 0$$

$$\mu_{POS_{R_b}(Q)}(2) = 0.8633$$

$$\mu_{POS_{R_b}(Q)}(3) = 0.4317$$

$$\mu_{POS_{R_b}(Q)}(4) = 0$$

$$\mu_{POS_{R_b}(Q)}(5) = 0$$

and

$$\mu_{POS_{R_b}(Q)}(6) = 0.8633$$

.

The resulting degree of dependency will be

$$\gamma'_{\{b\}}(\{q\}) = \frac{\Sigma_{x \in U} \mu_{POS_{R_b}(Q)}(x)}{|U|}$$

$$= 2.1583/6$$

$$= 0.3597$$

.

3.5 Conclusions

This chapter covered the preliminaries of rough set and fuzzy rough set along with an example explaining calculation procedure to obtain dependency measures for both methods. The analysis of RST and L-FRFS measures fuzzy rough set has been carried out to evaluate their suitability for the problems at hand for feature selection in which mainly numerical features (attributes) are there.

For using L-FRFS measures in the present work the motivations are as follows.

1. For the case of real valued features, for applying rough set method, we need to discretize the feature values. The discernibility of features are affected by the quantization and therefore becomes dependent on the quantization of feature values.

2. In case of fuzzy rough sets, the real feature values are taken as it is and therefore no such quantization is done and discernibility of features are therefore more accurate as well as meaningful.

3. For real valued features the number of quantization levels can be large or infinite. Thus to capture the feature values in a discrete sense is not simply possible.

4. If a real valued feature based problem is solved using rough set through discretization, it is highly possible that while application, a newer intermediate value for which the quantization level was not fixed may arise, making the rough set based method of no use since the nominations to values are predefined. In other words, new value is a new nomination and the rough set based feature reduction has to be done once again including the new nominations.

In the next chapter, a method of initialization has been proposed for the population based optimization methods and the same has been applied for feature selection using rough set and fuzzy rough set based dependency measure as fitness function.

Chapter 4

Improving the Initialization

4.1 Introduction

In the literature many authors have suggested different techniques for initialization of PSO. Qiang *et al.* [72] suggested chaotic initialization which uses something similar to pheromone used in ACO. Ruksaphil *et al.* [73] proposed minimax initialization which minimizes the maximum error. Paolo *et al.* [74] suggested alternate initialization technique of log, normal and lognormal distribution which replaces uniform distribution, Babu *et al.* [75] suggested two stage initialization, in which first stage is about selecting best strings, by evaluating these strings repeatedly with equal number of population size and in second stage these best strings get combined, and forms the new population, which is used for further operations. Guo *et al.* [76] suggested a re-initialization technique, which is based on estimation of the varieties and activities of the particles. In this method, group of particles which satisfies re-initialization pre-conditions, will be used for initialization, facilitating balance global search capabilities.

The above initialization methods are for general PSO algorithm; these are not specifically for feature selection. In the case of feature selection, an initialization method should ensure that distribution of strings in the population is uniform as far as a number of features selected are concerned. In this chapter a new initialization method based on this idea is developed and tested for feature selection problem.