Chapter 2

Rough Sets and Fuzzy-Rough Sets for Feature Selection

2.1 Introduction

Data reduction is a main point of interest across a wide variety of fields. In fact, focusing on this step is crucial as it often presents a source of significant data loss. Many techniques were proposed in literature to achieve the task of data reduction. However, most of them tend to destroy the underlying semantics of the features after reduction or require additional information about the given data set for thresholding. Thus, it seems necessary to think about a technique that can on the one hand reduce data dimensionality using information contained within the data set and on the other hand capable of preserving the meaning of the features. Rough Set Theory (RST) and Fuzzy-Rough Set Theory (FRST) can be used as such tools to discover data dependencies and to reduce the number of attributes contained in a data set using the data alone, requiring no additional information (Jensen, 2005).

This Chapter focuses on introducing RST and FRST for feature selection. In the next Section, a brief refresh on some tools for data reduction is proposed. In Section 2.3, the basic concepts of RST are highlighted and the fundamentals of FRST are presented next in Section 2.4.

2.2 Data Reduction Approaches

Dealing with high-dimensional data sets presents an urgent need for a set of tools to reduce data dimensionality. These techniques can be categorized into two heads: those that transform the underlying meaning of the features, called the *"transformation-based approaches"*, and those that are semantic-preserving techniques known as the *"selection-based approaches"*.

Transformation based approaches, also called *"feature extraction approaches"*, involve simplifying the amount of resources required to describe a large set of data accurately. Feature extraction is a general term for methods that construct combinations of variables to represent the original set of features but with new variables while still describing the data with sufficient accuracy. The transformation based techniques are employed in situations where the semantics of the original database are not needed by any future process.

In contrast to the semantics-destroying dimensionality reduction techniques, the semantics-preserving techniques, also called *"feature selection techniques"*, attempt to retain the meaning of the original feature set. The main aim of this kind of techniques is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features.

It is important to mention that data dimensionality reduction techniques suffer from some limitations. Most of these techniques involve the user for parameterizing the algorithms and this is a significant drawback. Some feature selectors require noise levels to be specified by the user beforehand, some simply rank features leaving the user to choose their own subset. There are those that require the user to state how many features are to be chosen, or they must supply a threshold that determines when the algorithm should terminate. All of these require the users to make a decision based on their own (possibly faulty) judgment (Jensen, 2005). To overcome the shortcomings of the existing methods, it would be interesting to look for a method that does not require any external or additional information to function appropriately. Rough Set Theory (RST) and Fuzzy-Rough Set Theory (FRST) can be used as such tools.

2.3 Rough Set Based Approach for Feature Selection

In this Section, rudiments of the RST will be outlined and basic concepts of the theory will be illustrated by a simple tutorial example.

2.3.1 Decision and Information Systems

Data are represented as a table where each row represents an object and where each column represents an attribute that can be measured for each object. Such table is called an *"Information System"* (IS) or an *"Information Table"* (IT). To fit this definition to our research field, an information table can be seen as a representation of antigens which are defined via a set of attributes. Formally, an information system can be defined as a pair $IS = (U, A)$ where $U = \{x_1, x_2, \ldots, x_n\}$ is a non-empty, finite set of objects called the *"universe"* and $A = \{a_1, a_2, \ldots, a_k\}$ is a non-empty, finite set of *"condition"* attributes. In supervised learning, a special case of the defined information table is considered, called a *"Decision Table"* (DT) or a *"Decision System"* (DS). A DT is an information system of the form $IS = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished attribute called *"decision"*. The value set of d, called $\theta = \{d_1, d_2, \ldots, d_s\}$.

Example 2.1 *Using the terminology of RST, the data set presented in Table 2.1 can be considered as an information system $IS = (U, A)$ consisting of 4 conditional features (a, b, c, d) and 8 objects. To illustrate an example of a decision system, Table 2.2 is used. It consists of 4 conditional features (a, b, c, d), 1 decision feature (e) and 8 objects.*

Table 2.1: Information System

$x \in U$	a	b	c	d
x_1	1	0	2	2
x_2	0	1	1	1
x_3	2	0	0	1
x_4	1	1	0	2
x_5	1	0	2	0
x_6	2	2	0	1
x_7	2	1	1	1
x_8	0	1	1	0

Table 2.2: Decision System

$x \in U$	a	b	c	d	e
x_1	1	0	2	2	0
x_2	0	1	1	1	2
x_3	2	0	0	1	1
x_4	1	1	0	2	2
x_5	1	0	2	0	1
x_6	2	2	0	1	1
x_7	2	1	1	1	2
x_8	0	1	1	0	1

$IND\_C = \{x_1\}, \ldots, \{x_8\}$

$\Rightarrow$ ils sont tous $\neq$

2.3.2 Indiscernibility Relation

The indiscernibility relation is the mathematical basis of rough set theory. Every object of the universe is described by certain amount of information expressed by means of some attributes used for object description. Objects characterized by the same information are indiscernible in view of the available information about them. For every set of attributes $P \subset A$, an indiscernibility relation, denoted by $IND(P)$ or U/P, is defined in the following way: two objects, x_i and x_j, are indiscernible by the set of attributes P in A, if $p(x_i) = p(x_j)$ for every $p \subset P$. In other words, two objects are considered to be indiscernible or equivalent if and only if they have the same values for all attributes in the set.

The equivalence class of $IND(P)$ is called elementary set in P because it represents the smallest discernible groups of objects. For any element x_i of U, the equivalence class of x_i in relation $IND(P)$ is represented as $[x_i]_{IND(P)}$. For every object $x_j \in U$, we will use $a_i(x_j)$ to denote the value of a condition attribute a_i for an object x_j. Similarly, $d(x_j)$ is the value of the decision attribute for an object x_j. We further extend these notations for a set of attributes $P \subseteq A$, by defining $P(x_j)$ to be value tuple of attributes in P for an object x_j. The indiscernibility relation based on a subset of the condition attributes P, denoted by $IND(P)$, is defined as follows:

$$IND(P) = U/P = \{[x_j]_P | x_j \in U\} \tag{2.1}$$

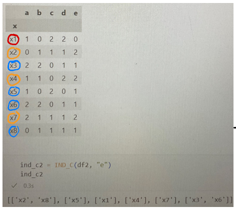
$$where \ \ [x_j]_P = \{x_i | P(x_i) = P(x_j)\} \tag{2.2}$$

The indiscernibility relation based on the decision attribute d, denoted by $IND_{\{d\}}$, is defined as follows:

$$IND_{\{d\}} = U/\{d\} = \{[x_j]_{\{d\}} | x_j \in U\} \tag{2.3}$$

Example 2.2 *In order to illustrate how a decision table from Table 2.2 defines an indiscernibility relation, we consider the following three non-empty subsets of the conditional attributes:* $\{a\}$, $\{b, c\}$ *and* $\{a, b, c\}$. *The relation* IND *may define three partitions of the universe.*

$$IND(a) = \{\{x_1, x_4, x_5\}, \{x_2, x_8\}, \{x_3, x_6, x_7\}\}$$
$$IND(b, c) = \{\{x_3\}, \{x_1, x_5\}, \{x_4\}, \{x_2, x_7, x_8\}, \{x_6\}\}$$



x	a	b	c	d	e
x1	1	0	2	2	0
x2	0	1	1	1	2
x3	2	2	0	1	1
x4	1	1	0	2	2
x5	1	0	2	0	1
x6	2	2	0	1	1
x7	2	1	1	1	2
x8	0	1	1	1	1

```
ind_c2 = IND_C(df2, "e")
ind_c2
```

✓ 0.3s

```
[['x2', 'x8'], ['x5'], ['x1'], ['x4'], ['x7'], ['x3', 'x6']]
```

$X_0 = \{o_j \mid e(o_j) = 0\}$

$= \{x_1\}$

$\underline{C} X_0 = \{x_1\}$

$X_1 = \{o_j \mid e(o_j) = 1\}$

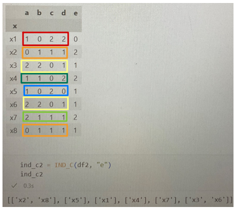
$= \{x_3, x_5, x_6, x_8\}$

$\underline{C} X_1 = \{x_3, x_6, x_5\}$

$X_2 = \{o_j \mid e(o_j) = 2\}$

$= \{x_2, x_4, x_7\}$

$\underline{C} X_2 = \{x_4, x_7\}$



x	a	b	c	d	e
x1	1	0	2	2	0
x2	0	1	1	1	2
x3	2	2	0	1	1
x4	1	1	0	2	2
x5	1	0	2	0	1
x6	2	2	0	1	1
x7	2	1	1	1	2
x8	0	1	1	1	1

```
ind_c2 = IND_C(df2, "e")
ind_c2
```

✓ 0.3s

```
[['x2', 'x8'], ['x5'], ['x1'], ['x4'], ['x7'], ['x3', 'x6']]
```

$$\Rightarrow POS_C\{(d)\} = \cup \underline{C} X$$

$$= \{x_1\} \cup \{x_3, x_6, x_5\} \cup \{x_4, x_7\}$$

$$POS_C\{(d)\} = \{x_1, x_3, x_4, x_5, x_6, x_7\}$$

OK

$$IND(a,b,c) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\}$$

If we take into consideration the set a, the objects x_1, x_4 and x_5 belong to the same equivalence class; they are indiscernible.

2.3.3 Lower and Upper Approximations

The rough set approach to data analysis hinges on two basic concepts, namely the lower and upper approximations of a set, referring to the elements that doubtless belong to the set, and to the elements that possibly belong to the set. Let $P \subseteq A$ and $X \subseteq U$. We can approximate X using only the information contained by constructing the P-lower and P-upper approximations of X, denoted $\underline{P}(X)$ and $\overline{P}(X)$ respectively where:

$$\underline{P}(X) = \{x|[x]_P \subseteq X\} \tag{2.4}$$

$$\overline{P}(X) = \{x|[x]_P \cap X \neq \emptyset\} \tag{2.5}$$

Objects in $\underline{P}(X)$ can be with certainty classified as members of X on the basis of knowledge in P, while objects in $\overline{P}(X)$ can be only classified as possible members of X on the basis of knowledge in P.

Let P and Q be equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{P}(X) \tag{2.6}$$

$$NEG_P(Q) = U - \bigcup_{X \in U/Q} \overline{P}(X) \tag{2.7}$$

$$BND_P(Q) = \bigcup_{X \in U/Q} \overline{P}(X) - \bigcup_{X \in U/Q} \underline{P}(X) \tag{2.8}$$

The positive region contains all objects of U that can be classified to classes of U/Q using the information in attributes P. The boundary region, $BND_P(Q)$, is the set of objects that can possibly, but not certainly, be classified in this way. The negative region, $NEG_P(Q)$, is the set of objects that cannot be classified to classes of U/Q.

Example 2.3 *Using the same decision table presented in Table 2.2, an illustrative example of the above mentioned calculations is given in what follows where $P = \{b, c\}$ and $Q = \{e\}$:*

$$\begin{aligned} POS_P(Q) &= \bigcup\{\emptyset, \{x_3, x_6\}, \{x_4\}\} = \{x_3, x_4, x_6\} \\ NEG_P(Q) &= U - \bigcup\{\{x_1, x_5\}, \{x_3, x_1, x_5, x_2, x_7, x_8, x_6\}, \{x_4, x_2, x_7, x_8\}\} = \emptyset \\ BND_P(Q) &= \bigcup\{\{x_1, x_5\}, \{x_3, x_1, x_5, x_2, x_7, x_8, x_6\}, \{x_4, x_2, x_7, x_8\}\} - \{x_3, x_4, x_6\} \\ &= \{x_1, x_2, x_5, x_7, x_8\} \end{aligned}$$

This means that objects x_3, x_4 and x_6 can certainly be classified as belonging to a class in attribute e, when considering attributes b and c. The rest of the objects cannot be classified as the information that would make them discernible is absent.

2.3.4 Independency of Attributes

An important issue in data analysis is discovering dependencies between attributes. Intuitively, a set of attributes Q depends totally on a set of attributes P, denoted $P \Rightarrow Q$, if all attribute values from Q are uniquely determined by values of attributes from P. If there exists a functional dependency between values of Q and P, then Q depends totally on P. In rough set theory, dependency is defined in the following way: For $P, Q \subset A$, it is said that Q depends on P in a degree $k (0 \leq k \leq 1)$, denoted $P \Rightarrow kQ$, if

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \tag{2.9}$$

If $k = 1$, Q depends totally on P, if $0 < k < 1$, Q depends partially (in a degree k) on P, and if $k = 0$ then Q does not depend on P.

By calculating the change in dependency when an attribute is removed from the set of considered conditional attributes, a measure of the significance of the attribute can be obtained. The higher the change in dependency, the more significant the attribute is. If the significance is 0, then the attribute is dispensable. More formally, given P, Q and an attribute $a \in P$:

$$\sigma_P(Q, a) = \gamma_P(Q) - \gamma_{P-\{a\}}(Q) \tag{2.10}$$

Example 2.4 *From Table 2.2, the dependency of attribute $\{e\}$ from the attributes $\{b, c\}$ is:*

$$\gamma_{\{b,c\}}(\{e\}) = \frac{|POS_{\{b,c\}}(\{e\})|}{|U|} = \frac{|\{x_3,x_4,x_6\}|}{|\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}|} = \frac{3}{8}$$

if $P = \{a, b, c\}$ and $Q = e$ then

$$\gamma_{\{a,b,c\}}(\{e\}) = \frac{|\{x_3,x_4,x_6,x_7\}}{8} = \frac{4}{8}\ ; \qquad \gamma_{\{a,b\}}(\{e\}) = \frac{|\{x_3,x_4,x_6,x_7\}}{8} = \frac{4}{8}$$
$$\gamma_{\{b,c\}}(\{e\}) = \frac{|\{x_3,x_4,x_6\}}{8} = \frac{3}{8}\ ; \qquad \gamma_{\{a,c\}}(\{e\}) = \frac{|\{x_3,x_4,x_6,x_7\}}{8} = \frac{4}{8}$$

And calculating the significance of the three attributes gives:

$$\sigma_P(Q, a) = \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{b,c\}}(\{e\}) = \frac{1}{8}$$
$$\sigma_P(Q, b) = \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,c\}}(\{e\}) = 0$$
$$\sigma_P(Q, c) = \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,b\}}(\{e\}) = 0$$

From this, it follows that attribute a is indispensable, but attributes b and c can be dispensed with when considering the dependency between the decision attribute and the given individual conditional attributes.

2.3.5 Core and Reduct of Attributes

The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision feature, D, as the original.

A subset $R \subseteq A$ is a reduct of A with respect to D, if R is independent and

$$\gamma_R(D) = \gamma_A(D) \tag{2.11}$$

Hence, a reduct is a set of attributes from A that preserves dependency and, consequently, set approximation. It means that a reduct is the minimal subset of attributes that enables the same classification of elements of the universe as the whole set of attributes.

The core is the intersection of all reducts. Hence, it is included in every reduct. In a sense, the core is the most important subset of attributes, for none of its elements can be removed without affecting the classification power of attributes.

Example 2.5 *Using the same decision table of Table 2.2, the dependencies for all possible subsets of A can be calculated:*

$$\gamma_{\{a,b,c,d\}}(\{e\}) = \tfrac{8}{8}\ ;\ \gamma_{\{a,b,c\}}(\{e\}) = \tfrac{4}{8}\ ;\ \gamma_{\{a,b,d\}}(\{e\}) = \tfrac{8}{8}\ ;\ \gamma_{\{a,c,d\}}(\{e\}) = \tfrac{8}{8}\ ;$$

$$\gamma_{\{b,c,d\}}(\{e\}) = \tfrac{8}{8}\ ;\ \gamma_{\{a,b\}}(\{e\}) = \tfrac{4}{8}\ ;\ \gamma_{\{a,c\}}(\{e\}) = \tfrac{4}{8}\ ;\ \gamma_{\{a,d\}}(\{e\}) = \tfrac{3}{8}\ ;$$

$$\gamma_{\{b,c\}}(\{e\}) = \tfrac{3}{8}\ ;\ \gamma_{\{b,d\}}(\{e\}) = \tfrac{8}{8}\ ;\ \gamma_{\{c,d\}}(\{e\}) = \tfrac{8}{8}\ ;\ \gamma_{\{a\}}(\{e\}) = \tfrac{0}{8}\ ;$$

$$\gamma_{\{b\}}(\{e\}) = \tfrac{1}{8}\ ;\ \gamma_{\{c\}}(\{e\}) = \tfrac{0}{8}\ ;\ \gamma_{\{d\}}(\{e\}) = \tfrac{2}{8}$$

Note that the given data set is consistent since $\gamma_{\{a,b,c,d\}}(\{e\}) = 1$. The minimal reduct set for this example is: Reduct = $\{\{b,d\},\{c,d\}\}$

Table 2.3: First Reduct

$x \in U$	b	d	e
x_1	0	2	0
x_2	1	1	2
x_3	0	1	1
x_4	1	2	2
x_5	0	0	1
x_6	2	1	1
x_7	1	1	2
x_8	1	0	1

Table 2.4: Second Reduct

$x \in U$	c	d	e
x_1	2	2	0
x_2	1	1	2
x_3	0	1	1
x_4	0	2	2
x_5	2	0	1
x_6	0	1	1
x_7	1	1	2
x_8	1	0	1

In Table 2.2, there are two possible reducts with respect to the decision attribute $\{e\}$; $\{b,d\}$ and $\{c,d\}$. These reducts are independent with respect to the decision attribute $\{e\}$ and have the same dependency as the whole subset of condition attributes A. That means that either the attribute b or c can be eliminated from the table and *consequently instead of Table 2.2, we can use either Table 2.3 or Table 2.4. The core is the attribute d. It is the intersection of the two possible reducts.*

2.3.6 Rough Sets and the Use of Heuristics

The problem of finding a reduct has been the subject of much research (Swiniarski & Skowron, 2003). The most basic solution to locating such a subset is to simply generate all possible subsets and retrieve those with a maximum rough set dependency degree. However, this is an expensive solution to the problem and is only practical for very simple data sets. Most of the time, only one reduct is required as, typically, only one subset of features is used to reduce a data set, so all the calculations involved in discovering the rest are pointless. Another shortcoming of finding all possible reducts using rough sets is to inquire about which is the best reduct for the classification process. The solution to these issues is to apply a *"heuristic attribute selection"* method (Zhong, Dong, & Ohsuga, 2001).

Among the most interesting heuristic methods proposed in literature, we mention the *"QuickReduct"* algorithm (Chouchoulas & Shen, 2001). QuickReduct attempts to calculate a reduct without exhaustively generating all possible subsets. It starts off with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough set dependency metric. According to the QuickReduct algorithm, the dependency of each attribute is calculated and the best candidate is chosen. This process continues until the dependency of the reduct equals the consistency of the data set (Jensen, 2005). Other approaches generating reducts from information systems have been developed and can be found in (Slezak, 1996).

2.3.7 Rough Sets and the Link to Other Theories

Rough set theory was combined with other theories such as the *probability theory*, *fuzzy set theory* (Zadeh, 1965) and the *belief function theory* (Shafer, 1976). As each of these theories has inherent characteristics, works tended to explore the effectiveness of the mentioned theories to handle specific kinds of problems. For instance, both rough sets and the probability theory were hybridized in order to tackle problems dealing with incomplete data, and to solve the same kind of problems rough set theory was combined with the belief function theory. Another combination of rough sets was its hybridization with fuzzy sets which gave rise to the Fuzzy-Rough Set Theory (FRST). The latter was, also, dedicated to handle incomplete data sets. Besides, it is applied to estimate the missing values in the learning process. FRST was first of all proposed as an extension to the rough set theory as the latter has some limitations. This will be discussed in the next Section.

2.4 Fuzzy-Rough Feature Selection

In this Section, the basics of FRST are described while focusing on the feature selection process performed by this theory.

2.4.1 From Rough Sets to Fuzzy-Rough Sets

In most databases, values of attributes may be both crisp and real-valued, and this is where many feature selectors including RST encounter a problem. More precisely, it is not possible to decide whether two attribute values are similar and to what extent they are the same. For instance, in RST, the values -0.1 and -0.11 are as different as -0.1 and 300. One solution to this problem is to discretize the data set beforehand producing a new database with crisp values. However, this is often still inadequate since it presents a kind of information loss (Jensen, 2005).

To deal with the issue stated above, it is clearly necessary to call for methods providing the means of feature selection for both crisp and real-value attributed data sets. One theory which is capable to handle such type of data is fuzzy set theory (Zadeh, 1965). An introduction to fuzzy set theory can be found in Appendix A. Fuzzy sets and the process of fuzzification provide a mechanism by which real-valued features can be effectively managed. This is achieved by allowing values to belong to more than one label with various degrees of membership. Consequently, the vagueness presented in data can be modeled. And therefore, the hybridization of RST and fuzzy set theory, giving rise to the fuzzy-rough set theory, is seen as a promising combination to be applied for the feature selection task. In fact, FRST is an extension of rough set theory allowing all memberships to take values in the range $[0, 1]$. This permits a higher degree of flexibility compared to the strict requirements of rough sets that only deal with full or zero set memberships (Jensen, 2005).

2.4.2 Fuzzy Equivalence Classes

In the same way that crisp equivalence classes are central to rough sets, fuzzy equivalence classes are central to the fuzzy-rough set approach. For typical rough set attribute reduction applications, this means that the decision values and the conditional values may all be fuzzy. The concept of crisp equivalence classes can be extended by the inclusion of a fuzzy similarity relation S on the universe, which determines the extent to which two elements are similar in S. For example, if $\mu_S(x, y) = 0.9$, then objects x and y are considered to be quite similar. The usual properties of reflexivity ($\mu_S(x, x) = 1$), symmetry ($\mu_S(x, y) = \mu_S(y, x)$) and transitivity ($\mu_S(x, z) \geq \mu(x, y) \wedge \mu_S(y, z)$) hold (Jensen, 2005). Using the fuzzy similarity relation, the fuzzy equivalence class $[x]_S$ for objects close to x can be defined:

$$\mu_{[x]_S}(y) = \mu_S(x, y) \tag{2.12}$$

This definition degenerates to the normal definition of equivalence classes when S is non-fuzzy. The following axioms should hold for a fuzzy equivalence class F:

1. $\exists x, \mu_F(x) = 1$ (μ_F is normalized)
2. $\mu_F(x) \wedge \mu_S(x, y) \leq \mu_F(y)$ class of y
3. $\mu_F(x) \wedge \mu_F(y) \leq \mu_S(x, y)$

The first axiom corresponds to the requirement that an equivalence class is nonempty. The second axiom states that elements in y's neighborhood are in the equivalence class of y. The final axiom states that any two elements in F are related via S. The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes.

Example 2.6 *Consider the crisp partitioning of a universe of discourse, U, by the attributes in Q* : $U/Q = \{\{x_1, x_3, x_6\}, \{x_2, x_4, x_5\}\}$. *This contains two equivalence classes* ($\{x_1, x_3, x_6\}$ *and* $\{x_2, x_4, x_5\}$) *that can be thought of as degenerated fuzzy sets, with those elements belonging to the class possessing a membership of one, zero otherwise. For the first class, for instance, the objects* x_2, x_4 *and* x_5 *have a membership of zero. Extending this to the case of fuzzy equivalence classes is straightforward: objects can be allowed to assume membership values, with respect to any given class, in the interval* [0, 1].

2.4.3 Fuzzy-Rough Sets Basic Concepts

Same as RST, FRST is based on the fuzzy P-lower and fuzzy P-upper approximations which are defined as:

$$\mu_{\underline{P}X}(F_i) = inf_x max\{1 - \mu_{F_i}(x), \mu_X(x)\}\forall i \tag{2.13}$$

$$\mu_{\overline{P}X}(F_i) = sup_x min\{\mu_{F_i}(x), \mu_X(x)\}\forall i \tag{2.14}$$

where F_i denotes a fuzzy equivalence class belonging to U/P, and X is the fuzzy concept to be approximated. The tuple $< \underline{P}X, \overline{P}X >$ is called a *fuzzy-rough set*. These definitions diverge a little from the crisp upper and lower approximations, as the memberships of individual objects to the approximations are not explicitly available. As a result of this, the fuzzy lower and upper approximations are herein redefined as:

$$\mu_{\underline{P}X} = sup_{F\in U/P} min(\{\mu_F(x), inf_{y\in U} max\{1 - \mu_F(y), \mu_X(y)\}) \tag{2.15}$$

$$\mu_{\overline{P}X} = sup_{F\in U/P} min(\{\mu_F(x), sup_{y\in U} min\{\mu_F(y), \mu_X(y)\}) \tag{2.16}$$

2.4.4 Fuzzy-Rough Reduction Process

The fuzzy-rough set based feature selection approach was built on the notion of fuzzy lower approximation to enable reduction of data sets containing real-valued features. As will be shown, the process becomes identical to RST when dealing with nominal well-defined features. The crisp positive region in traditional rough set theory is defined as the union of the lower approximations. By the extension principle, the membership of an object $x \in U$, belonging to the fuzzy positive region can be defined by:

$$\mu_{POS_{P(Q)}}(x) = sup_{X\in U/Q} \mu_{\underline{P}X}(x) \tag{2.17}$$

Object x will not belong to the positive region only if the equivalence class it belongs to is not a constituent of the positive region. This is equivalent to RST where objects belong to the positive region only if their underlying equivalence class does so. Similarly, the negative and boundary regions can be defined. Using the definition of the fuzzy positive region, the new dependency function can be defined as follows:

$$\gamma'_P(Q) = \frac{|\mu_{POS_{P(Q)}}(x)|}{|U|} = \frac{|\sum_{x\in U} \mu_{POS_{P(Q)}}(x)|}{|U|} \tag{2.18}$$

As with rough sets, the dependency of Q on P is the proportion of objects that are discernible out of the entire data set. In the present approach, this corresponds to determining the fuzzy cardinality of $\mu_{POS_{P(Q)}}(x)$ divided by the total number of objects in the universe. The definition of dependency degree covers the crisp case as its specific instance.

2.4.5 Fuzzy-Rough QuickReduct

A problem may arise when this approach is compared to RST. In conventional rough set attribute reduction, a reduct is defined as a subset R of the features which have the same information content as the full feature set A. In terms of the dependency function this means that the values $\gamma(R)$ and $\gamma(A)$ are identical and equal to 1 if the data set is consistent. However, in the fuzzy-rough approach this is not necessarily the case as the uncertainty encountered when objects belong to many fuzzy equivalence classes results in a reduced total dependency. With these issues in mind, a fuzzy-rough QuickReduct algorithm, described by Algorithm 2.1, was developed in (Jensen, 2005).

Algorithm 2.1 The Fuzzy-Rough QuickReduct Algorithm

```
C: the set of all conditional features;
D: the set of decision features;
R ← { } ; $\gamma'_{best} = 0$ ; $\gamma'_{prev} = 0$
do
    T ← R
    $\gamma'_{prev} = \gamma'_{best}$
    $\forall x \in (C - R)$
        if $\gamma'_{R\cup\{x\}}(D) > \gamma'_{T}(D)$
            T ← $R \cup \{x\}$
            $\gamma'_{best} = \gamma'_{T}(D)$
    R ← T
until $\gamma'_{best} = \gamma'_{prev}$
return R
```

The fuzzy-rough QuickReduct algorithm employs the fuzzy-rough dependency function γ' to choose those attributes to add to the current reduct candidate. The algorithm terminates when the addition of any remaining attribute does not increase the dependency.

Example 2.6 *To illustrate the operation of the fuzzy-rough QuickReduct algorithm, an example database is given in Table 2.5. This Table contains 3 real-valued conditional attributes and 1 crisp-valued decision attribute.*

Table 2.5: Example Database

Object	a	b	c	q
x_1	-0.4	-0.3	-0.5	No
x_2	-0.4	0.2	-0.1	Yes
x_3	-0.3	-0.4	-0.3	No
x_4	0.3	-0.3	0	Yes
x_5	0.2	-0.3	0	Yes
x_6	0.2	0	0	No

Let us remind that in rough set feature selection, the data set would be discretized using the non-fuzzy sets. However, in the fuzzy-rough approach fuzzy sets, defined in Figure 2.1, are used in calculating the fuzzy lower ap-

proximations and fuzzy positive regions. This will be illustrated in the following example.

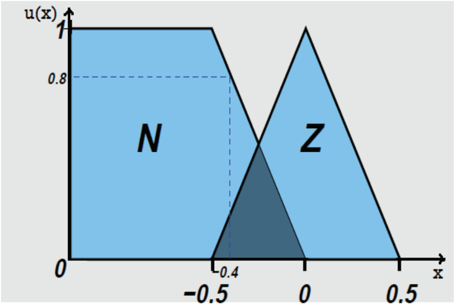


Figure 2.1: Fuzzification for Conditional Features

To start with, and by using the fuzzy-rough QuickReduct algorithm, the potential reduct is initialized to the empty set.

Using fuzzy sets which are defined in Figure 2.1 and setting $A = \{a\}$, $B = \{b\}$, $C = \{c\}$ *and* $Q = \{q\}$, *the following equivalence classes are obtained:*

$$
\begin{aligned}
U/A &= \{N_a, Z_a\}\\
U/B &= \{N_b, Z_b\}\\
U/C &= \{N_c, Z_c\}\\
U/Q &= \{\{x_1, x_3, x_6\}, \{x_2, x_4, x_5\}\}
\end{aligned}
$$

The first step is to calculate the lower approximations of the sets A, B and C, using Equation 2.15. To clarify the calculations involved, Table 2.6 contains the membership degrees of objects to fuzzy equivalence classes. For simplicity, only A will be considered here; i.e, using A to approximate Q.

Table 2.6: Membership Values of Objects to Corresponding Fuzzy Sets

Object	a		b		c		q	
	N_a	Z_a	N_b	Z_b	N_c	Z_c	$\{x_1, x_3, x_6\}$	$\{x_2, x_4, x_5\}$
x_1	0.8	0.2	0.6	0.4	1.0	0.0	1.0	0.0
x_2	0.8	0.2	0.0	0.6	0.2	0.8	0.0	1.0
x_3	0.6	0.4	0.8	0.2	0.6	0.4	1.0	0.0
x_4	0.0	0.4	0.6	0.4	0.0	1.0	0.0	1.0
x_5	0.0	0.6	0.6	0.4	0.0	1.0	0.0	1.0
x_6	0.0	0.6	0.0	1.0	0.0	1.0	1.0	0.0

For the first decision equivalence class $X = \{x_1, x_3, x_6\}, \mu_{\underline{A}\{x_1,x_3,x_6\}}(x)$ *needs to be calculated:*

$$\mu_{\underline{A}\{x_1,x_3,x_6\}}(x) = sup_{F\in U/A}min(\{\mu_F(x), inf_{y\in U}max\{1-\mu_F(y), \mu_{\{x_1,x_3,x_6\}}(y)\}\})$$

Considering the first fuzzy equivalence class of A, N_a:

$$min(\mu_{N_a}(x), inf_{y\in U}max\{1-\mu_{N_a}(y), \mu_{\{x_1,x_3,x_6\}}(y)\})$$

For object x_2 this can be calculated as follows: From Table 2.6, it can be seen that the membership of object x_2 to the fuzzy equivalence class N_a, $\mu_{N_a}(x_2)$, is 0.8. The remainder of the calculation involves finding the smallest of the following values:

$$\begin{aligned}
max(1-\mu_{N_a}(x_1), \mu_{\{x_1,x_3,x_6\}}(x_1)) &= max(0.2, 1.0) = 1.0\\
max(1-\mu_{N_a}(x_2), \mu_{\{x_1,x_3,x_6\}}(x_2)) &= max(0.2, 0.0) = 0.2\\
max(1-\mu_{N_a}(x_3), \mu_{\{x_1,x_3,x_6\}}(x_3)) &= max(0.4, 1.0) = 1.0\\
max(1-\mu_{N_a}(x_4), \mu_{\{x_1,x_3,x_6\}}(x_4)) &= max(1.0, 0.0) = 1.0\\
max(1-\mu_{N_a}(x_5), \mu_{\{x_1,x_3,x_6\}}(x_5)) &= max(1.0, 0.0) = 1.0\\
max(1-\mu_{N_a}(x_6), \mu_{\{x_1,x_3,x_6\}}(x_6)) &= max(1.0, 1.0) = 1.0
\end{aligned}$$

From the calculations above, the smallest value is 0.2, hence:

$$min(\mu_{N_a}(x), inf_{y\in U}max\{1-\mu_{N_a}(y), \mu_{\{x_1,x_3,x_6\}}(y)\}) = min(0.8, inf\{1, 0.2, 1, 1, 1, 1\}) = 0.2$$

Similarly for Z_a:

$$min(\mu_{Z_a}(x), inf_{y\in U}max\{1-\mu_{Z_a}(y), \mu_{\{x_1,x_3,x_6\}}(y)\}) = min(0.2, inf\{1, 0.8, 1, 0.6, 0.4, 1\}) = 0.2$$

Thus,

$$\mu_{\underline{A}\{x_1,x_3,x_6\}}(x_2) = 0.2$$

Calculating the A-lower approximation of $X = \{x_1, x_3, x_6\}$ for every object gives:

$$\begin{aligned}
\mu_{\underline{A}\{x_1,x_3,x_6\}}(x_1) = 0.2\ ;\ \mu_{\underline{A}\{x_1,x_3,x_6\}}(x_2) = 0.2\\
\mu_{\underline{A}\{x_1,x_3,x_6\}}(x_3) = 0.4\ ;\ \mu_{\underline{A}\{x_1,x_3,x_6\}}(x_4) = 0.4\\
\mu_{\underline{A}\{x_1,x_3,x_6\}}(x_5) = 0.4\ ;\ \mu_{\underline{A}\{x_1,x_3,x_6\}}(x_6) = 0.4
\end{aligned}$$

The corresponding values for $X = \{x_2, x_4, x_5\}$ can, also, be determined:

$$\begin{aligned}
\mu_{\underline{A}\{x_2,x_4,x_5\}}(x_1) = 0.2\ ;\ \mu_{\underline{A}\{x_2,x_4,x_5\}}(x_2) = 0.2\\
\mu_{\underline{A}\{x_2,x_4,x_5\}}(x_3) = 0.4\ ;\ \mu_{\underline{A}\{x_2,x_4,x_5\}}(x_4) = 0.4\\
\mu_{\underline{A}\{x_2,x_4,x_5\}}(x_5) = 0.4\ ;\ \mu_{\underline{A}\{x_2,x_4,x_5\}}(x_6) = 0.4
\end{aligned}$$

It is a coincidence here that $\mu_{\underline{A}\{x_2,x_4,x_5\}}(x) = \mu_{\underline{A}\{x_1,x_3,x_6\}}(x)$ *for this example. Using these values, the fuzzy positive region for each object can be calculated using Equation 2.17:*

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

This results in:

$$\begin{aligned} \mu_{POS_{A(Q)}}(x_1) = 0.2 \ ; \ \mu_{POS_{A(Q)}}(x_2) = 0.2 \\ \mu_{POS_{A(Q)}}(x_3) = 0.4 \ ; \ \mu_{POS_{A(Q)}}(x_4) = 0.4 \\ \mu_{POS_{A(Q)}}(x_5) = 0.4 \ ; \ \mu_{POS_{A(Q)}}(x_6) = 0.4 \end{aligned}$$

The next step is to determine the degree of dependency of Q on A using Equation 2.18:

$$\gamma'_A(Q) = \frac{|\sum_{x \in U} \mu_{POS_{A}(Q)}(x)|}{|U|} = \frac{2}{6}$$

Calculating for B and C gives:

$$\gamma'_B(Q) = \frac{2.4}{6} \ ; \ \gamma'_C(Q) = \frac{1.6}{6}$$

From this, it can be seen that attribute b will cause the greatest increase in dependency degree. This attribute is chosen and added to the potential reduct. The process iterates and the two dependency degrees calculated are:

$$\gamma'_{\{a,b\}}(Q) = \frac{3.4}{6} \ ; \ \gamma'_{\{b,c\}}(Q) = \frac{3.2}{6}$$

Adding attribute a to the reduct candidate causes the larger increase of dependency, so the new candidate becomes $\{a, b\}$*. Lastly, attribute c is added to the potential reduct:*

$$\gamma'_{\{a,b,c\}}(Q) = \frac{3.4}{6}$$

As this causes no increase in dependency, the algorithm stops and outputs the reduct $\{a, b\}$*. The data set can now be reduced to only those attributes appearing in the reduct. When rough set attribute reduction is performed on this data set (after using the same fuzzy sets to discretize the real-valued attributes), the reduct generated is* $\{a, b, c\}$*; i.e. the full conditional attribute set (Jensen & Shen, 2002). Unlike rough set attribute reduction, the true minimal reduct was found using the information on degrees of membership. It is clear from this example that the information lost by using rough set attribute reduction can be important when trying to discover the smallest reduct from a data set.*

2.5 Conclusion

In this Chapter, both rough sets and fuzzy-rough sets for feature selection were introduced. The application of these theories form the main contribution of this Thesis. New hybridized models will be presented in the second part of this dissertation where the mentioned theories are combined with the DCA. This is to cover the DCA limitations which are linked to its data pre-processing step.