New approaches to fuzzy-rough feature selection

Jensen, Richard; Shen, Qiang

Published in:
IEEE Transactions on Fuzzy Systems

DOI:
10.1109/TFUZZ.2008.924209

Publication date:
2009

Citation for published version (APA):
Jensen, R., & Shen, Q. (2009). New approaches to fuzzy-rough feature selection. *IEEE Transactions on Fuzzy Systems*, *17*(4), 824-838. https://doi.org/10.1109/TFUZZ.2008.924209

New Approaches to Fuzzy-Rough Feature Selection

Richard Jensen and Qiang Shen

***Abstract*— There has been great interest in developing methodologies which are capable of dealing with imprecision and uncertainty. The large amount of research currently being carried out in fuzzy and rough sets is representative of this. Many deep relationships have been established and recent studies have concluded at the complementary nature of the two methodologies. Therefore, it is desirable to extend and hybridize the underlying concepts to deal with additional aspects of data imperfection. Such developments offer a high degree of flexibility and provide robust solutions and advanced tools for data analysis. Fuzzy-rough set-based feature selection has been shown to be highly useful at reducing data dimensionality, but possesses several problems that render it ineffective for large datasets. This paper proposes three new approaches to fuzzy-rough feature selection based on fuzzy similarity relations. In particular, a fuzzy extension to crisp discernibility matrices is proposed and utilized. Initial experimentation shows that the methods greatly reduce dimensionality whilst preserving classification accuracy.**

***Index Terms*— Dimensionality reduction; feature selection; fuzzy-rough sets; fuzzy discernibility matrix; fuzzy positive region; fuzzy boundary region.**

I. Introduction

FEATURE selection (FS) [7], [15] addresses the problem of selecting those input features that are most predictive of a given outcome; a problem encountered in many areas of computational intelligence. Unlike other dimensionality reduction methods, feature selectors preserve the original meaning of the features after reduction. This has found application in tasks that involve datasets containing huge numbers of features (in the order of tens of thousands) which, for some learning algorithms, might be impossible to process further. Recent examples include text processing and web content classification [13].

There are often many features involved, and combinatorially large numbers of feature combinations, to select from. Note that the number of feature subset combinations with m features from a collection of N total features is $N!/[m!(N-m)!]$. It might be expected that the inclusion of an increasing number of features would increase the likelihood of including enough information to distinguish between classes. Unfortunately, this is not necessarily true if the size of the training dataset does not also increase rapidly with each additional feature included. A high-dimensional dataset increases the chances that a learning algorithm will find spurious patterns that are not valid in general. More features may introduce more measurement noise, and hence reduce performance (e.g. classification accuracy). Most techniques employ some degree of reduction in order to cope with large amounts of data, so an efficient and effective reduction method is required.

R. Jensen and Q. Shen are with the Department of Computer Science, The University of Wales, Aberystwyth, Ceredigion, SY23 3DB, Wales, UK. Email: {rkj,qqs}@aber.ac.uk

Lately there has been great interest in developing methodologies which are capable of dealing with imprecision and uncertainty, and the resounding amount of research currently being done in the areas related to fuzzy [43] and rough sets [20] is representative of this. The success of rough set theory is due in part to three aspects of the theory. Firstly, only the facts hidden in data are analysed. Secondly, no additional information about the data is required for data analysis such as thresholds or expert knowledge on a particular domain. Thirdly, it finds a minimal knowledge representation for data. As rough set theory handles only one type of imperfection found in data, it is complementary to other concepts for the purpose, such as fuzzy set theory. The two fields may be considered analogous in the sense that both can tolerate inconsistency and uncertainty - the difference being the type of uncertainty and their approach to it; fuzzy sets are concerned with vagueness, rough sets are concerned with indiscernibility. Many deep relationships have been established and more so, most of the recent studies have concluded at this complementary nature of the two methodologies, especially in the context of granular computing. Therefore, it is desirable to extend and hybridize the underlying concepts to deal with additional aspects of data imperfection. Such developments offer a high degree of flexibility and provide robust solutions and advanced tools for data analysis [16].

Fuzzy-rough feature selection (FRFS) provides a means by which discrete or real-valued noisy data (or a mixture of both) can be effectively reduced without the need for user-supplied information. Additionally, this technique can be applied to data with continuous or nominal decision attributes, and as such can be applied to regression as well as classification datasets. The only additional information required is in the form of fuzzy partitions for each feature which can be automatically derived from the data. However, there are several problems with the approach from theoretical and practical viewpoints that motivate further developments in this area. This paper proposes three new methods for fuzzy-rough feature selection that address these problems and provide robust strategies for dimensionality reduction. In particular, the notion of the fuzzy discernibility matrix is proposed for computing reductions.

This paper is structured as follows. The theoretical background is given in section II, providing necessary details for crisp rough set theory, discernibility matrices and fuzzy-rough concepts. In the third section, the new developments for fuzzy-rough feature selection are presented: fuzzy lower approximation-based, fuzzy boundary region-based and fuzzy discernibility matrix-based approaches are discussed. Some initial experimentation is provided in section IV. The paper is concluded in section V.

II. Theoretical Background

Rough Set Attribute Reduction (RSAR) [5] provides a filter-based tool by which knowledge may be extracted from a domain in a concise way; retaining the information content whilst reducing the amount of knowledge involved. The main advantage that rough set analysis has is that it requires no additional parameters to operate other than the supplied data [10]. It works by making use of the granularity structure of the data only. This is a major difference when compared with Dempster-Shafer theory [25] and fuzzy set theory which require probability assignments and membership values respectively. However, this does not mean that *no* model assumptions are made. In fact by using only the given information, the theory assumes that the data is a true and accurate reflection of the real world (which may not be the case). The numerical and other contextual aspects of the data are ignored which may seem to be a significant omission, but keeps model assumptions to a minimum.

An example dataset is given in table I to illustrate the concepts involved. Here, the table consists of four conditional features (a, b, c, d), one decision feature (e) and eight objects.

TABLE I
AN EXAMPLE DATASET

$x \in \mathbb{U}$	a	b	c	d	$\Rightarrow$	e
0	S	R	T	T		R
1	R	S	S	S		T
2	T	R	R	S		S
3	S	S	R	T		T
4	S	R	T	R		S
5	T	T	R	S		S
6	T	S	S	S		T
7	R	S	S	R		S

A. Rough Set Feature Selection

Central to RSAR is the concept of indiscernibility. Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe of discourse) and $\mathbb{A}$ is a non-empty finite set of attributes such that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that attribute a may take. With any $P \subseteq \mathbb{A}$ there is an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in \mathbb{U}^2 | \forall a \in P,\ a(x) = a(y)\} \quad (1)$$

The partition of $\mathbb{U}$, generated by *IND(P)* is denoted $\mathbb{U}$*/IND(P)* (or $\mathbb{U}$*/P* for simplicity) and can be calculated as follows:

$$\mathbb{U}/IND(P) = \otimes\{\mathbb{U}/IND(\{a\}) | a \in P\}, \quad (2)$$

where $\otimes$ is specifically defined as follows for sets A and B:

$$A \otimes B = \{X \cap Y | X \in A, Y \in B, X \cap Y \neq \emptyset\} \quad (3)$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_P$. For the illustrative example, if *P* = {*b,c*}, then objects 1, 6 and 7 are indiscernible; as are objects 0 and 4. *IND(P)* creates the following partition of $\mathbb{U}$:

$$\begin{aligned} \mathbb{U}/IND(P) &= \mathbb{U}/IND(\{b\}) \otimes \mathbb{U}/IND(\{c\}) \\ &= \{\{0,2,4\},\{1,3,6,7\},\{5\}\} \otimes \\ &\quad \{\{2,3,5\},\{1,6,7\},\{0,4\}\} \\ &= \{\{2\},\{0,4\},\{3\},\{1,6,7\},\{5\}\} \end{aligned}$$

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

$$\underline{P}X = \{x \in \mathbb{U} \mid [x]_P \subseteq X\} \quad (4)$$

$$\overline{P}X = \{x \in \mathbb{U} \mid [x]_P \cap X \neq \emptyset\} \quad (5)$$

The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called a rough set. Let *P* and *Q* be sets of attributes inducing equivalence relations over $\mathbb{U}$, then the positive, negative and boundary regions can be defined as:

$$\begin{aligned} POS_P(Q) &= \textstyle\bigcup_{X \in \mathbb{U}/Q} \underline{P}X \\ NEG_P(Q) &= \mathbb{U} - \textstyle\bigcup_{X \in \mathbb{U}/Q} \overline{P}X \\ BND_P(Q) &= \textstyle\bigcup_{X \in \mathbb{U}/Q} \overline{P}X - \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \end{aligned}$$

The positive region contains all objects of $\mathbb{U}$ that can be classified to classes of $\mathbb{U}$*/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 $\mathbb{U}$*/Q*. For example, let *P* = {*b,c*} and *Q* = {*e*}, then

$$\begin{aligned} POS_P(Q) &= \textstyle\bigcup\{\emptyset, \{2,5\}, \{3\}\} = \{2,3,5\} \\ NEG_P(Q) &= \mathbb{U} - \textstyle\bigcup\{\{0,4\}, \{2,0,4,1,6,7,5\}, \{3,1,6,7\}\} \\ &= \emptyset \\ BND_P(Q) &= \mathbb{U} - \{2,3,5\} = \{0,1,4,6,7\} \end{aligned}$$

This means that objects 2, 3 and 5 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.

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* $\subset \mathbb{A}$, it is said that *Q* depends on *P* in a degree *k* ($0 \leq k \leq 1$), denoted *P* $\Rightarrow_k$ *Q*, if

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|} \quad (6)$$

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. In the example, the degree of dependency of attribute {*e*} on the attributes {*b,c*} is:

$$\gamma_{\{b,c\}}(\{e\}) = \frac{|POS_{\{b,c\}}(\{e\})|}{|\mathbb{U}|}$$

$$= \frac{|\{2,3,5\}|}{|\{0,1,2,3,4,5,6,7\}|} = \frac{3}{8}$$

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 $a \in P$,

$$\sigma_P(Q,a) = \gamma_P(Q) - \gamma_{P-\{a\}}(Q) \quad (7)$$

1) Reduction Method: 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 attribute as the original. A *reduct* R_{min} is defined as a minimal subset *R* of the initial attribute set $\mathbb{C}$ such that for a given set of attributes D, $\gamma_R(\mathbb{D}) = \gamma_{\mathbb{C}}(\mathbb{D})$. From the literature, R is a minimal subset if $\gamma_{R-\{a\}}(\mathbb{D}) \neq \gamma_R(\mathbb{D})$ for all $a \in R$. This means that no attributes can be removed from the subset without affecting the dependency degree. Hence, a minimal subset by this definition may not be the *global* minimum (a reduct of smallest cardinality). A given dataset may have many reduct sets, and the collection of all reducts is denoted by

$$\begin{aligned} R_{all} = \{X \mid X \subseteq \mathbb{C}, \gamma_X(\mathbb{D}) = \gamma_{\mathbb{C}}(\mathbb{D}); \\ \gamma_{X-\{a\}}(\mathbb{D}) \neq \gamma_X(\mathbb{D}),\ \forall a \in X\} \end{aligned} \quad (8)$$

The intersection of all the sets in R_{all} is called the *core*, the elements of which are those attributes that cannot be eliminated without introducing more contradictions to the representation of the dataset. For many tasks (for example, feature selection [7]), a reduct of minimal cardinality is ideally searched for. That is, an attempt is to be made to locate a single element of the reduct set $R_{min} \subseteq R_{all}$:

$$R_{min} = \{X \mid X \in R_{all},\ \forall Y \in R_{all}, |X| \leq |Y|\} \quad (9)$$

The intersection of all the reducts is called the *core*, the elements of which are those attributes that cannot be eliminated without introducing more contradictions to the dataset. The goal of RSAR is to discover reducts.

Using the example, the dependencies for all possible subsets of $\mathbb{C}$ can be calculated:

$$\begin{array}{ll} \gamma_{\{a,b,c,d\}}(\{e\}) = 8/8 & \gamma_{\{b,c\}}(\{e\}) = 3/8 \\ \gamma_{\{a,b,c\}}(\{e\}) = 4/8 & \gamma_{\{b,d\}}(\{e\}) = 8/8 \\ \gamma_{\{a,b,d\}}(\{e\}) = 8/8 & \gamma_{\{c,d\}}(\{e\}) = 8/8 \\ \gamma_{\{a,c,d\}}(\{e\}) = 8/8 & \gamma_{\{a\}}(\{e\}) = 0/8 \\ \gamma_{\{b,c,d\}}(\{e\}) = 8/8 & \gamma_{\{b\}}(\{e\}) = 1/8 \\ \gamma_{\{a,b\}}(\{e\}) = 4/8 & \gamma_{\{c\}}(\{e\}) = 0/8 \\ \gamma_{\{a,c\}}(\{e\}) = 4/8 & \gamma_{\{d\}}(\{e\}) = 2/8 \\ \gamma_{\{a,d\}}(\{e\}) = 3/8 & \end{array}$$

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

The problem of finding a reduct of an information system has been the subject of much research [1], [28]. The QUICKREDUCT algorithm given in figure 1 (adapted from [5]), attempts to calculate reducts 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, until this produces its maximum possible value for the dataset. The heuristic used is based on equation (7), where $\sigma_{P\cup a}(Q,a)$ is evaluated for each attribute, given reduct candidate P. Other such techniques may be found in [21], [22].

QUICKREDUCT($\mathbb{C}$,$\mathbb{D}$).
$\mathbb{C}$, the set of all conditional attributes;
$\mathbb{D}$, the set of decision attributes.

(1) $R \leftarrow \{\}$
(2) **do**
(3) $\quad T \leftarrow R$
(4) $\quad$ **foreach** $x \in (\mathbb{C} - R)$
(5) $\quad\quad$ **if** $\gamma_{R\cup\{x\}}(\mathbb{D}) > \gamma_T(\mathbb{D})$
(6) $\quad\quad\quad T \leftarrow R \cup \{x\}$
(7) $\quad R \leftarrow T$
(8) **until** $\gamma_R(\mathbb{D}) == \gamma_{\mathbb{C}}(\mathbb{D})$
(9) **return** R

Fig. 1. The QUICKREDUCT Algorithm

According to the QUICKREDUCT algorithm, the dependency degree of the addition of each attribute to the current reduct candidate (initially empty) is calculated, and the best candidate chosen. This process continues until the dependency of the subset equals the consistency of the dataset (1 if the dataset is consistent). The generated reduct shows the way of reducing the dimensionality of the original dataset by eliminating those conditional attributes that do not appear in the set.

Determining the consistency of the entire dataset is reasonable for many datasets. However, it may be infeasible for very large data, so alternative stopping criteria may have to be used. One such criterion could be to terminate the search when there is no further increase in the dependency measure [5].

This, however, is not guaranteed to find a true reduct, i.e. one that is of minimal cardinality. Using the dependency function to discriminate between candidates may lead the search down a non-minimal path. It is impossible to predict which combinations of attributes will lead to an optimal reduct based on changes in dependency with the addition or deletion of single attributes. It does result in a close-to-minimal subset, though, which is still useful in greatly reducing dataset dimensionality.

B. Discernibility Matrix Approach

Many applications of rough sets to feature selection make use of discernibility matrices for finding reducts. A discernibility matrix [14], [26] of a decision table $D = (\mathbb{U}, \mathbb{C} \cup \mathbb{D})$ is a symmetric $|\mathbb{U}| \times |\mathbb{U}|$ matrix with entries defined:

$$c_{ij} = \{a \in \mathbb{C} | a(x_i) \neq a(x_j)\} \quad i,j = 1,...,|\mathbb{U}| \quad (10)$$

Each c_{ij} contains those attributes that differ between objects i and j. For finding reducts, the decision-relative discernibility

matrix is of more interest. This only considers those object discernibilities that occur when the corresponding decision features differ. Returning to the example dataset, the decision-relative discernibility matrix found in Table II is produced. For example, it can be seen from Table I that objects 0 and 1 differ in each attribute. Although some attributes in objects 1 and 3 differ, their corresponding decisions are the same so no entry appears in the decision-relative matrix. Grouping all entries containing single features forms the core of the dataset (those features appearing in *every* reduct). Such entries imply that at least two objects can only be distinguished by this feature alone, and so must appear in all reducts. Here, the core of the dataset is $\{d\}$.

From this, the discernibility function can be defined. This is a concise notation of how each object within the dataset may be distinguished from the others. A discernibility function f_D is a boolean function of m boolean variables $a_1^*, ..., a_m^*$ (corresponding to the attributes $a_1, ..., a_m$) defined as below:

$$f_D(a_1^*, ..., a_m^*) = \wedge\{\vee c_{ij}^* | 1 \leq j \leq i \leq |\mathbb{U}|, c_{ij} \neq \emptyset\} \quad (11)$$

where $c_{ij}^* = \{a^* | a \in c_{ij}\}$. By finding the set of all prime implicants [26] of the discernibility function, all the minimal reducts of a system may be determined.

From table II, the decision-relative discernibility function is (with duplicates removed):

$$\begin{aligned} f_D(a,b,c,d) = \quad & \{a \vee b \vee c \vee d\} \wedge \{a \vee c \vee d\} \wedge \{b \vee c\} \\ & \wedge \{d\} \wedge \{a \vee b \vee c\} \wedge \{a \vee b \vee d\} \\ & \wedge \{b \vee c \vee d\} \wedge \{a \vee d\} \end{aligned}$$

Further simplification can be performed by removing those sets that are supersets of others:

$$f_D(a,b,c,d) = \quad \{b \vee c\} \wedge \{d\}$$

The reducts of the dataset may be obtained by converting the above expression from conjunctive normal form to disjunctive normal form (without negations). Hence, the minimal reducts are $\{b, d\}$ and $\{c, d\}$. Although this is guaranteed to discover all minimal subsets, it is a costly operation rendering the method impractical for even medium-sized datasets.

For certain applications, a single minimal subset is all that is required for data reduction. For example, dimensionality reduction within text classification tends to use only one subset to remove unnecessary keywords [11]. This has led to approaches that consider finding individual shortest prime implicants from the discernibility function. A common method is to incrementally add those attributes that occur with the most frequency in the function, removing any clauses containing the attributes, until all clauses are eliminated [17], [32]. However, even this does not ensure that a minimal subset is found - the search can proceed down non-minimal paths.

C. Fuzzy-Rough Feature Selection

The RSAR process described previously can only operate effectively with datasets containing discrete values. Additionally, there is no way of handling noisy data. As most datasets contain real-valued attributes, it is necessary to perform a discretization step beforehand. This is typically implemented by standard fuzzification techniques [24], enabling linguistic labels to be associated with attribute values. It also aids the modelling of uncertainty in data by allowing the possibility of the membership of a value to more than one linguistic label. However, membership degrees of attribute values to fuzzy sets are not exploited in the process of dimensionality reduction. By using *fuzzy-rough* sets [9], [19], it is possible to use this information to better guide feature selection [13].

1) 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 [9], [29], [39]. For typical 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. 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 partitioning of the universe of discourse can play the role of fuzzy equivalence classes [9]. Consider the crisp partitioning of a universe of discourse, $\mathbb{U}$, by the attributes in Q: $\mathbb{U}/Q$ = {{1,3,6},{2,4,5}}. This contains two equivalence classes ({1,3,6} and {2,4,5}) 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 2, 4 and 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]. $\mathbb{U}/Q$ is not restricted to crisp partitions only; fuzzy partitions are equally acceptable. For the work presented here, a simple fuzzification pre-processor is used to derive the fuzzy sets, corresponding to fuzzy equivalence classes, via the use of the statistical properties of the data.

2) Fuzzy-Rough Sets: There have been two main lines of thought in the hybridization of fuzzy and rough sets, the constructive approach and the axiomatic approach. A general framework for the study of fuzzy-rough sets from both of these viewpoints is presented in [42]. For the constructive approach, generalized lower and upper approximations are defined based on fuzzy relations. Initially, these were fuzzy similarity/equivalence relations [9] but have since been extended to arbitrary fuzzy relations [42]. The axiomatic approach is primarily for the study of the mathematical properties of fuzzy-rough sets [36]. Here, various classes of fuzzy-rough approximation operators are characterized by different sets of axioms that guarantee the existence of types of fuzzy relations producing the same operators.

An original definition for fuzzy P-lower and P-upper approximations was given as follows [9]:

$$\mu_{\underline{P}X}(F_i) = \inf_x \max\{1 - \mu_{F_i}(x), \mu_X(x)\} \quad \forall i \quad (12)$$

$$\mu_{\overline{P}X}(F_i) = \sup_x \min\{\mu_{F_i}(x), \mu_X(x)\} \quad \forall i \quad (13)$$

where F_i is a fuzzy equivalence class and X is the (fuzzy) concept to be approximated. The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called a fuzzy-rough set. These definitions diverge a little from the crisp upper and lower approximations, as the memberships

TABLE II
THE DECISION-RELATIVE DISCERNIBILITY MATRIX

	0	1	2	3	4	5	6	7
0								
1	$\{a,b,c,d\}$							
2	$\{a,c,d\}$	$\{a,b,c\}$						
3	$\{b,c\}$		$\{a,b,d\}$					
4	$\{d\}$	$\{a,b,c,d\}$		$\{b,c,d\}$				
5	$\{a,b,c,d\}$	$\{a,b,c\}$		$\{a,b,d\}$				
6	$\{a,b,c,d\}$		$\{b,c\}$		$\{a,b,c,d\}$	$\{b,c\}$		
7	$\{a,b,c,d\}$	$\{d\}$		$\{a,c,d\}$			$\{a,d\}$	

of individual objects to the approximations are not explicitly available. As a result of this, the fuzzy lower and upper approximations are redefined as [12]:

$$\mu_{\underline{P}X}(x) = \sup_{F\in\mathbb{U}/P} \min(\mu_F(x), \inf_{y\in\mathbb{U}} \max\{1-\mu_F(y), \mu_X(y)\}) \tag{14}$$

$$\mu_{\overline{P}X}(x) = \sup_{F\in\mathbb{U}/P} \min(\mu_F(x), \sup_{y\in\mathbb{U}} \min\{\mu_F(y), \mu_X(y)\}) \tag{15}$$

It can be seen that these definitions degenerate to traditional rough sets when all equivalence classes are crisp [11].

Also defined in the literature are rough-fuzzy sets [9], which can be seen to be a particular case of fuzzy-rough sets. A rough-fuzzy set is a generalization of a rough set, derived from the approximation of a fuzzy set in a crisp approximation space. In [38] it is argued that, to be consistent, the approximation of a crisp set in a fuzzy approximation space should be called a fuzzy-rough set, and the approximation of a fuzzy set in a crisp approximation space should be called a rough-fuzzy set, making the two models complementary. In this framework, the approximation of a fuzzy set in a fuzzy approximation space is considered to be a more general model, unifying the two theories. However, most researchers consider the traditional definition of fuzzy-rough sets in [9] as standard.

The specific use of $\min$ and $\max$ operators in the definitions above is expanded in [23], where a broad family of fuzzy-rough sets is constructed, each member represented by a particular implicator and t-norm. The properties of three well-known implicators (S-, R- and QL-implicators) are investigated. Further investigations in this area can be found in [8], [29], [37], [42].

3) Fuzzy-Rough Reduction Process: Fuzzy-rough set-based feature selection builds on the notion of the fuzzy lower approximation to enable reduction of datasets containing real-valued attributes. As will be shown, the process becomes identical to the crisp approach when dealing with nominal well-defined attributes.

The crisp positive region in traditional rough set theory is defined as the union of the lower approximations. By the extension principle [44], the membership of an object $x \in \mathbb{U}$, belonging to the fuzzy positive region can be defined by

$$\mu_{POS_P(Q)}(x) = \sup_{X\in\mathbb{U}/Q} \mu_{\underline{P}X}(x) \tag{16}$$

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 the crisp version where objects belong to the positive region only if their underlying equivalence class does so.

Using the definition of the fuzzy positive region, the fuzzy-rough dependency function can be defined as follows:

$$\gamma'_P(Q) = \frac{|\mu_{POS_P(Q)}(x)|}{|\mathbb{U}|} = \frac{\sum_{x\in\mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|} \tag{17}$$

As with crisp rough sets, the dependency of Q on P is the proportion of objects that are discernible out of the entire dataset. 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.

If the fuzzy-rough reduction process is to be useful, it must be able to deal with multiple attributes, finding the dependency between various subsets of the original attribute set. For example, it may be necessary to be able to determine the degree of dependency of the decision attribute(s) with respect to P = $\{a, b\}$. In the crisp case, $\mathbb{U}/P$ contains sets of objects grouped together that are indiscernible according to both attributes a and b. In the fuzzy case, objects may belong to many equivalence classes, so the cartesian product of $\mathbb{U}/IND(\{a\})$ and $\mathbb{U}/IND(\{b\})$ must be considered in determining $\mathbb{U}/P$. In general,

$$\mathbb{U}/P = \otimes\{\mathbb{U}/IND(\{a\}) | a \in P\}, \tag{18}$$

where

$$A \otimes B = \{X \cap Y | X \in A, Y \in B, X \cap Y \neq \emptyset\} \tag{19}$$

Each set in $\mathbb{U}/P$ denotes an equivalence class. For example, if P = $\{a, b\}$, $\mathbb{U}/IND(\{a\})$ = $\{N_a, Z_a\}$ and $\mathbb{U}/IND(\{b\})$ = $\{N_b, Z_b\}$, then

$$\mathbb{U}/P = \{N_a \cap N_b, N_a \cap Z_b, Z_a \cap N_b, Z_a \cap Z_b\}$$

The extent to which an object belongs to such an equivalence class is therefore calculated by using the conjunction of constituent fuzzy equivalence classes, say F_i, $i = 1, 2, ..., n$:

$$\mu_{F_1\cap...\cap F_n}(x) = \min(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)) \tag{20}$$

4) Fuzzy-Rough QUICKREDUCT*:* A problem may arise when this approach is compared to the crisp approach. In conventional RSAR, a reduct is defined as a subset R of the attributes which have the same information content as the full attribute 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 dataset 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.

FRQUICKREDUCT($\mathbb{C}$,$\mathbb{D}$).
$\mathbb{C}$, the set of all conditional attributes;
$\mathbb{D}$, the set of decision attributes.

(1) $R \leftarrow \{\}$; $\gamma'_{best} = 0$; $\gamma'_{prev} = 0$
(2) **do**
(3) $T \leftarrow R$
(4) $\gamma'_{prev} = \gamma'_{best}$
(5) **foreach** $x \in (\mathbb{C} - R)$
(6) **if** $\gamma'_{R\cup\{x\}}(\mathbb{D}) > \gamma'_T(\mathbb{D})$
(7) $T \leftarrow R \cup \{x\}$
(8) $\gamma'_{best} = \gamma'_T(\mathbb{D})$
(9) $R \leftarrow T$
(10) **until** $\gamma'_{best} == \gamma'_{prev}$
(11) **return** R

Fig. 2. The fuzzy-rough QUICKREDUCT algorithm

Object	a	b	c	q
1	-0.4	-0.3	-0.5	no
2	-0.4	0.2	-0.1	yes
3	-0.3	-0.4	-0.3	no
4	0.3	-0.3	0	yes
5	0.2	-0.3	0	yes
6	0.2	0	0	no

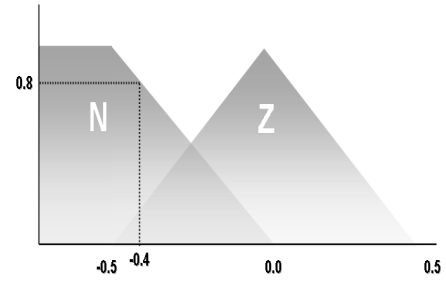

Fig. 3. Dataset and corresponding fuzzy sets

With these issues in mind, a fuzzy-rough hill-climbing search algorithm has been developed as given in Fig. 2. It employs the fuzzy-rough dependency function γ' to choose which attributes to add to the current reduct candidate in a manner similar to QUICKREDUCT. The algorithm terminates when the addition of any remaining attribute does not increase the dependency (such a criterion could be used with the QUICKREDUCT algorithm). As this fuzzy-rough degree of dependency measure is non-monotonic, it is possible that the hill-climbing search terminates having reached only a local optimum. The global optimum may lie elsewhere in the search space. As with the original QUICKREDUCT algorithm, the algorithm may return a super-reduct (i.e. a reduct containing superfluous features) due to the non-optimality of the search heuristic used [40].

Note that with the fuzzy-rough QUICKREDUCT algorithm, for a dimensionality of n, $(n^2+n)/2$ evaluations of the dependency function may be performed for the worst-case dataset. However, as FRFS is used for dimensionality reduction prior to any involvement of the system which will employ those attributes belonging to the resultant reduct, this operation has no negative impact upon the run-time efficiency of the system.

5) Example: To illustrate the operation of FRFS, an example dataset is given in Fig. 3. In crisp RSAR, the dataset would be discretized using non-fuzzy sets. However, in the new approach membership degrees are used in calculating the fuzzy lower approximations and fuzzy positive regions. To begin with, the fuzzy-rough QUICKREDUCT algorithm initializes the potential reduct (i.e. the current best set of attributes) to the empty set.

Using the fuzzy sets defined in Fig. 3 (for all conditional attributes for illustrative simplicity), and setting $A = \{a\}$, $B = \{b\}$, $C = \{c\}$ and $Q = \{q\}$, the following equivalence classes are obtained:

$$\begin{aligned} \mathbb{U}/A &= \{N_a, Z_a\} \\ \mathbb{U}/B &= \{N_b, Z_b\} \\ \mathbb{U}/C &= \{N_c, Z_c\} \\ \mathbb{U}/Q &= \{\{1,3,6\},\{2,4,5\}\} \end{aligned}$$

The first step is to calculate the lower approximations of the decision concepts for the sets A, B and C. For straightforwardness, only the calculations involving A are demonstrated here; that is, using A to approximate Q. For the first decision equivalence class X = {1,3,6}, $\mu_{\underline{A}\{1,3,6\}}(x)$ is calculated:

$$\mu_{\underline{A}\{1,3,6\}}(x) = \sup_{F\in\mathbb{U}/A} \min(\mu_F(x), \inf_{y\in\mathbb{U}} \max\{1-\mu_F(y), \mu_{\{1,3,6\}}(y)\})$$

Considering the first fuzzy equivalence class of A, N_a:

$$\min(\mu_{N_a}(x), \inf_{y\in\mathbb{U}} \max\{1-\mu_{N_a}(y), \mu_{\{1,3,6\}}(y)\})$$

For object 2 this can be calculated as follows:

$$\min(0.8, \inf\{1, 0.2, 1, 1, 1, 1\}) = 0.2$$

Similarly for Z_a

$$\min(0.2, \inf\{1, 0.8, 1, 0.6, 0.4, 1\} = 0.2$$

Thus,

$$\mu_{\underline{A}\{1,3,6\}}(2) = 0.2$$

Calculating the A-lower approximation of $X = \{1,3,6\}$ for every object gives

$$\begin{aligned} \mu_{\underline{A}\{1,3,6\}}(1) &= 0.2 \quad \mu_{\underline{A}\{1,3,6\}}(2) = 0.2 \\ \mu_{\underline{A}\{1,3,6\}}(3) &= 0.4 \quad \mu_{\underline{A}\{1,3,6\}}(4) = 0.4 \\ \mu_{\underline{A}\{1,3,6\}}(5) &= 0.4 \quad \mu_{\underline{A}\{1,3,6\}}(6) = 0.4 \end{aligned}$$

The corresponding values for X = $\{2,4,5\}$ can also be determined this way. Using these values, the fuzzy positive region for each object can be calculated via using

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

This results in:

$$\begin{aligned} \mu_{POS_A(Q)}(1) &= 0.2 \quad \mu_{POS_A(Q)}(2) = 0.2 \\ \mu_{POS_A(Q)}(3) &= 0.4 \quad \mu_{POS_A(Q)}(4) = 0.4 \\ \mu_{POS_A(Q)}(5) &= 0.4 \quad \mu_{POS_A(Q)}(6) = 0.4 \end{aligned}$$

It is a coincidence here that $\mu_{POS_A(Q)}(x) = \mu_{\underline{A}\{1,3,6\}}(x)$ for this example. The next step is to determine the degree of dependency of Q on A:

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

Similarly, 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 dataset can now be reduced to only those attributes appearing in the reduct. When crisp RSAR is performed on this dataset (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.

D. Problems with FRFS

FRFS has been shown to be a highly useful technique in reducing data dimensionality [13]. However, several problems exist with the method. Firstly, the complexity of calculating the Cartesian product of fuzzy equivalence classes becomes prohibitively high for large feature subsets. If the number of fuzzy sets per attribute is n, $n^{|R|}$ equivalence classes must be considered per attribute for feature subset R. Optimizations that attempt to alleviate this problem are given in [2], [13], but the complexity is still too high. In [3], a compact computational domain is proposed to reduce the computational effort required to calculate fuzzy lower approximations for large datasets, based on some of the properties of fuzzy connectives.

Secondly, it was shown in [30] that in some situations, the fuzzy lower approximation might not be a subset of the fuzzy upper approximation. This is undesirable from a theoretical viewpoint as it is meaningless for a lower approximation of a concept to be larger than its upper approximation as this suggests that there is more certainty in the upper than the lower. It was also shown that the Cartesian product of fuzzy equivalence classes might not result in a family of fuzzy equivalence classes. These issues motivate the development of the techniques proposed in this paper.

III. New Fuzzy Rough Feature Selection

This section presents three new techniques for fuzzy-rough feature selection, based on fuzzy similarity relations.

A. Fuzzy Lower Approximation-based FS

The previous method for fuzzy-rough feature selection used a fuzzy partitioning of the input space in order to determine fuzzy equivalence classes. Alternative definitions for the fuzzy lower and upper approximations can be found in [23], where a T-transitive fuzzy similarity relation is used to approximate a fuzzy concept X:

$$\mu_{\underline{R_P}X}(x) = \inf_{y\in\mathbb{U}} I(\mu_{R_P}(x,y), \mu_X(y)) \tag{21}$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y\in\mathbb{U}} T(\mu_{R_P}(x,y), \mu_X(y)) \tag{22}$$

Here, I is a fuzzy implicator and T a t-norm. R_P is the fuzzy similarity relation induced by the subset of features P:

$$\mu_{R_P}(x,y) = \bigcap_{a\in P} \{\mu_{R_a}(x,y)\} \tag{23}$$

$\mu_{R_a}(x,y)$ is the degree to which objects x and y are similar for feature a. Many fuzzy similarity relations can be constructed for this purpose, for example:

$$\mu_{R_a}(x,y) = 1 - \frac{|a(x) - a(y)|}{|a_{max} - a_{min}|} \tag{24}$$

$$\mu_{R_a}(x,y) = \exp(-\frac{(a(x) - a(y))^2}{2{\sigma_a}^2}) \tag{25}$$

$$\mu_{R_a}(x,y) = \max(\min(\frac{(a(y) - (a(x) - \sigma_a))}{(a(x) - (a(x) - \sigma_a))}, \frac{((a(x) + \sigma_a) - a(y))}{((a(x) + \sigma_a) - a(x))}, 0) \tag{26}$$

where ${\sigma_a}^2$ is the variance of feature a. As these relations do not necessarily display T-transitivity, the fuzzy transitive closure must be computed for each attribute [8]. The combination of feature relations in equation (23) has been shown to preserve T-transitivity [31].

1) Reduction: In a similar way to the original FRFS approach, the fuzzy positive region can be defined as:

$$\mu_{POS_{R_P}(Q)}(x) = \sup_{X\in\mathbb{U}/Q} \mu_{\underline{R_P}X}(x) \tag{27}$$

The resulting degree of dependency is:

$$\gamma'_P(Q) = \frac{\sum_{x\in\mathbb{U}} \mu_{POS_{R_P}(Q)}(x)}{|\mathbb{U}|} \tag{28}$$

A fuzzy-rough reduct R can be defined as a subset of features that preserves the dependency degree of the entire dataset, i.e. $\gamma'_R(\mathbb{D}) = \gamma'_{\mathbb{C}}(\mathbb{D})$. Based on this, a new fuzzy-rough QuickReduct algorithm can be constructed that operates in the same way as Fig. 2, but uses equation (28) to gauge subset quality. A proof of the monotonicity of the dependency function can be found in the appendix. Core features may be determined by considering the change in dependency of the full set of conditional features when individual attributes are removed:

$$Core(\mathbb{C}) = \{a \in \mathbb{C} | \gamma'_{\mathbb{C}-\{a\}}(Q) < \gamma'_{\mathbb{C}}(Q)\} \tag{29}$$

2) Example: The fuzzy connectives chosen for this example (and all others in this section) are the Łukasiewicz t-norm ($\max(x+y-1,0)$) and the Łukasiewicz fuzzy implicator ($\min(1-x+y,1)$). As recommended in [8], the Łukasiewicz t-norm is used as this produces fuzzy T-equivalence relations dual to that of a pseudo-metric. The use of the Łukasiewicz fuzzy implicator is also recommended as it is both a residual and S-implicator.

Using the fuzzy similarity measure defined in (26), the resulting relations are as follows for each feature in the dataset:

$$R_a(x,y) = \begin{pmatrix} 1.0 & 1.0 & 0.699 & 0.0 & 0.0 & 0.0 \\ 1.0 & 1.0 & 0.699 & 0.0 & 0.0 & 0.0 \\ 0.699 & 0.699 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.699 & 0.699 \\ 0.0 & 0.0 & 0.0 & 0.699 & 1.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 0.699 & 1.0 & 1.0 \end{pmatrix}$$

$$R_b(x,y) = \begin{pmatrix} 1.0 & 0.0 & 0.568 & 1.0 & 1.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.137 \\ 0.568 & 0.0 & 1.0 & 0.568 & 0.568 & 0.0 \\ 1.0 & 0.0 & 0.568 & 1.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 0.568 & 1.0 & 1.0 & 0.0 \\ 0.0 & 0.137 & 0.0 & 0.0 & 0.0 & 1.0 \end{pmatrix}$$

$$R_c(x,y) = \begin{pmatrix} 1.0 & 0.0 & 0.036 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.036 & 0.518 & 0.518 & 0.518 \\ 0.036 & 0.036 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.518 & 0.0 & 1.0 & 1.0 & 1.0 \\ 0.0 & 0.518 & 0.0 & 1.0 & 1.0 & 1.0 \\ 0.0 & 0.518 & 0.0 & 1.0 & 1.0 & 1.0 \end{pmatrix}$$

Again, the first step is to compute the lower approximations of each concept for each feature. Considering feature a and the decision concept $\{1,3,6\}$ in the example dataset:

$$\mu_{\underline{R_a}\{1,3,6\}}(x) = \inf_{y\in\mathbb{U}} I(\mu_{R_a}(x,y), \mu_{\{1,3,6\}}(y))$$

For object 3, this is

$$\begin{aligned} \mu_{\underline{R_a}\{1,3,6\}}(3) &= \inf_{y\in\mathbb{U}} I(\mu_{R_a}(3,y), \mu_{\{1,3,6\}}(y)) \\ &= \inf\{I(0.699,1), I(0.699,0), I(1,1), \\ &\qquad I(0,0), I(0,0), I(0,1)\} \\ &= 0.301 \end{aligned}$$

For the remaining objects, this is:

$$\begin{aligned} \mu_{\underline{R_a}\{1,3,6\}}(1) &= 0.0 \\ \mu_{\underline{R_a}\{1,3,6\}}(2) &= 0.0 \\ \mu_{\underline{R_a}\{1,3,6\}}(4) &= 0.0 \\ \mu_{\underline{R_a}\{1,3,6\}}(5) &= 0.0 \\ \mu_{\underline{R_a}\{1,3,6\}}(6) &= 0.0 \end{aligned}$$

For concept $\{2,4,5\}$, the lower approximations are:

$$\begin{aligned} \mu_{\underline{R_a}\{2,4,5\}}(1) &= 0.0 \\ \mu_{\underline{R_a}\{2,4,5\}}(2) &= 0.0 \\ \mu_{\underline{R_a}\{2,4,5\}}(3) &= 0.0 \\ \mu_{\underline{R_a}\{2,4,5\}}(4) &= 0.301 \\ \mu_{\underline{R_a}\{2,4,5\}}(5) &= 0.0 \\ \mu_{\underline{R_a}\{2,4,5\}}(6) &= 0.0 \end{aligned}$$

Hence, the positive regions for each object are:

$$\begin{aligned} \mu_{POS_{R_a}(Q)}(1) &= 0.0 \\ \mu_{POS_{R_a}(Q)}(2) &= 0.0 \\ \mu_{POS_{R_a}(Q)}(3) &= 0.301 \\ \mu_{POS_{R_a}(Q)}(4) &= 0.301 \\ \mu_{POS_{R_a}(Q)}(5) &= 0.0 \\ \mu_{POS_{R_a}(Q)}(6) &= 0.0 \end{aligned}$$

The resulting degree of dependency is therefore:

$$\begin{aligned} \gamma'_{\{a\}}(Q) &= \frac{\sum_{x\in\mathbb{U}} \mu_{POS_{R_a}(Q)}(x)}{|\mathbb{U}|} \\ &= \frac{0.602}{6} \\ &= 0.1003 \end{aligned}$$

Calculating the dependency degrees for the remaining features results in

$$\gamma'_{\{b\}}(Q) = 0.3597 \quad \gamma'_{\{c\}}(Q) = 0.4078$$

As feature c results in the largest increase in dependency degree, this feature is selected and added to the reduct candidate. The algorithm then evaluates the addition of all remaining features to this candidate. Fuzzy similarity relations are combined using (23). This produces the following evaluations:

$$\gamma'_{\{a,c\}}(Q) = 0.5501 \quad \gamma'_{\{b,c\}}(Q) = 1.0$$

Feature subset $\{b,c\}$ produces the maximum dependency value for this dataset, and the algorithm terminates. The dataset can now be reduced to these features only. The complexity of the algorithm is the same as that of FRFS in terms of the number of dependency evaluations. However, the explosive growth of the number of considered fuzzy equivalence classes is avoided through the use of fuzzy similarity relations and (23). This ensures that for one subset, only one fuzzy similarity relation is used to compute the fuzzy lower approximation.

B. Fuzzy Boundary Region-based FS

Most approaches to crisp rough set FS and all approaches to fuzzy-rough FS use only the lower approximation for the evaluation of feature subsets. The lower approximation contains information regarding the extent of certainty of object membership to a given concept. However, the upper approximation contains information regarding the degree of uncertainty of objects and hence this information can be used to discriminate between subsets. For example, two subsets may

result in the same lower approximation but one subset may produce a smaller upper approximation. This subset will be more useful as there is less uncertainty concerning objects within the boundary region (the difference between upper and lower approximations). The fuzzy-rough boundary region for a fuzzy concept X may thus be defined:

$$\mu_{BND_{R_P}(X)}(x) = \mu_{\overline{R_P}X}(x) - \mu_{\underline{R_P}X}(x) \tag{30}$$

The fuzzy-rough negative region for all decision concepts can be defined as follows:

$$\mu_{NEG_{R_P}}(x) = N(\sup_{X \in \mathbb{U}/Q} \mu_{\overline{R_P}X}(x)) \tag{31}$$

In classical rough set theory, the negative region is always empty for partitions [41]. It is interesting to note that the fuzzy-rough negative region is also always empty when the decisions are crisp. However, this is not necessarily the case when decisions are fuzzy. Further details can be found in the appendix.

1) Reduction: As the search for an optimal subset progresses, the object memberships to the boundary region for each concept diminishes until a minimum is achieved. For crisp rough set FS, the boundary region will be zero for each concept when a reduct is found. This may not necessarily be the case for fuzzy-rough FS due to the additional uncertainty involved. The uncertainty for a concept X using features in P can be calculated as follows:

$$U_P(X) = \frac{\sum\limits_{x \in \mathbb{U}} \mu_{BND_{R_P}(X)}(x)}{|\mathbb{U}|} \tag{32}$$

This is the average extent to which objects belong to the fuzzy boundary region for the concept X. The total uncertainty degree for all concepts, given a feature subset P is defined as:

$$\lambda_P(Q) = \frac{\sum\limits_{X \in \mathbb{U}/Q} U_P(X)}{|\mathbb{U}/Q|} \tag{33}$$

This is related to the conditional entropy measure which considers a combination of conditional probabilities $H(Q|P)$ in order to gauge the uncertainty present using features in P. In the crisp case, the minimization of this measure can be used to discover reducts: if the entropy for a feature subset P is zero, then the subset is a reduct [12].

Again, a QUICKREDUCT-style algorithm can be constructed for locating fuzzy-rough reducts based on this measure. Instead of maximising the dependency degree, the task of the algorithm is to minimize the total uncertainty degree. When this reaches the minimum for the dataset, a fuzzy-rough reduct has been found. A proof of the monotonicity of the total uncertainty degree can be found in the appendix.

2) Example: To determine the fuzzy boundary region, the lower and upper approximations of each concept for each feature must be calculated. Considering feature a and concept $\{1,3,6\}$:

$$\mu_{BND_{R_a}(\{1,3,6\})}(x) = \mu_{\overline{R_a}\{1,3,6\}}(x) - \mu_{\underline{R_a}\{1,3,6\}}(x)$$

For object 4, this is

$$\begin{aligned}\mu_{BND_{R_a}(\{1,3,6\})}(4) &= \sup_{y \in \mathbb{U}} T(\mu_{R_a}(4,y), \mu_{\{1,3,6\}}(y)) \\ &\quad - \inf_{y \in \mathbb{U}} I(\mu_{R_a}(4,y), \mu_{\{1,3,6\}}(y)) \\ &= 0.699 - 0.0 \\ &= 0.699\end{aligned}$$

For the remaining objects, this is:

$$\begin{aligned}\mu_{BND_{R_a}(\{1,3,6\})}(1) &= 1.0 \\ \mu_{BND_{R_a}(\{1,3,6\})}(2) &= 1.0 \\ \mu_{BND_{R_a}(\{1,3,6\})}(3) &= 0.699 \\ \mu_{BND_{R_a}(\{1,3,6\})}(5) &= 1.0 \\ \mu_{BND_{R_a}(\{1,3,6\})}(6) &= 1.0\end{aligned}$$

Hence, the uncertainty for concept $\{1,3,6\}$ is:

$$\begin{aligned}U_a(\{1,3,6\}) &= \frac{\sum\limits_{x \in \mathbb{U}} \mu_{BND_{R_a}(\{1,3,6\})}(x)}{|\mathbb{U}|} \\ &= \frac{1.0 + 1.0 + 0.699 + 0.699 + 1.0 + 1.0}{6} \\ &= 0.899\end{aligned}$$

For concept $\{2,4,5\}$, the uncertainty is:

$$\begin{aligned}U_a(\{2,4,5\}) &= \frac{\sum\limits_{x \in \mathbb{U}} \mu_{BND_{R_a}(\{2,4,5\})}(x)}{|\mathbb{U}|} \\ &= \frac{1.0 + 1.0 + 0.699 + 0.699 + 1.0 + 1.0}{6} \\ &= 0.899\end{aligned}$$

From this, the total uncertainty for feature a is calculated as follows:

$$\begin{aligned}\lambda_a(Q) &= \frac{\sum\limits_{X \in \mathbb{U}/Q} U_a(X)}{|\mathbb{U}/Q|} \\ &= \frac{0.899 + 0.899}{2} \\ &= 0.899\end{aligned} \tag{34}$$

The values of the total uncertainty for the remaining features are:

$$\lambda_{\{b\}}(Q) = 0.640 \quad \lambda_{\{c\}}(Q) = 0.592$$

As feature c results in the smallest total uncertainty, it is chosen and added to the reduct candidate. The algorithm then considers the addition of the remaining features to the subset:

$$\lambda_{\{a,c\}}(Q) = 0.500 \quad \lambda_{\{b,c\}}(Q) = 0.0$$

The subset $\{b,c\}$ results in the minimal uncertainty for the dataset, and the algorithm terminates. This is the same subset as that chosen by the fuzzy lower approximation-based method above. Again, the complexity of the algorithm is the same as that of FRFS, but avoids the Cartesian product of fuzzy equivalence classes. However, for each evaluation, both the fuzzy lower and upper approximations are considered and hence the calculation of the fuzzy boundary region is more costly than that of the fuzzy lower approximation alone.

C. Fuzzy Discernibility Matrix-based FS

As mentioned previously, there are two main branches of research in crisp rough set-based FS: those based on the dependency degree and those based on discernibility matrices. The developments given above are solely concerned with the extension of the dependency degree to the fuzzy-rough case. Hence, methods constructed based on the crisp dependency degree can be employed for fuzzy-rough FS.

By extending the discernibility matrix to the fuzzy case, it is possible to employ approaches similar to those in crisp rough set FS to determine fuzzy-rough reducts. A first step toward this is presented in [30], [33] where a crisp discernibility matrix is constructed for fuzzy-rough selection. A threshold is used, breaking the rough set ideology, which determines which features are to appear in the matrix entries. However, information is lost in this process as membership degrees are not considered. Search based on the crisp discernibility may result in reducts that are not true fuzzy-rough reducts.

1) Fuzzy Discernibility: The approach presented here extends the crisp discernibility matrix by employing fuzzy clauses. Each entry in the fuzzy discernibility matrix is a fuzzy set, to which every feature belongs to a certain degree. The extent to which a feature a belongs to the fuzzy clause C_{ij} is determined by the fuzzy discernibility measure:

$$\mu_{C_{ij}}(a) = N(\mu_{R_a}(i,j)) \tag{35}$$

where N denotes fuzzy negation and $\mu_{R_a}(i,j)$ is the fuzzy similarity of objects i and j, and hence $\mu_{C_{ij}}(a)$ is a measure of the fuzzy discernibility. For the crisp case, if $\mu_{C_{ij}}(a)$ = 1 then the two objects are distinct for this feature; if $\mu_{C_{ij}}(a)$ = 0, the two objects are identical. For fuzzy cases where $\mu_{C_{ij}}(a) \in (0,1)$, the objects are partly discernible. (The choice of fuzzy similarity relation must be identical to that of the fuzzy-rough dependency degree approach to find corresponding reducts.) Each entry in the fuzzy indiscernibility matrix is then a set of attributes and their corresponding memberships:

$$C_{ij} = \{a_x | a \in \mathbb{C}, x = N(\mu_{R_a}(i,j))\} \;\; i,j = 1,...,|\mathbb{U}| \tag{36}$$

For example, an entry C_{ij} in the fuzzy discernibility matrix might be:

$$C_{ij} : \{a_{0.4}, b_{0.8}, c_{0.2}, d_{0.0}\}$$

This denotes that $\mu_{C_{ij}}(a) = 0.4$, $\mu_{C_{ij}}(b) = 0.8$, etc. In crisp discernibility matrices, these values are either 0 or 1 as the underlying relation is an equivalence relation. The example clause can be viewed as indicating the value of each feature - the extent to which the feature discriminates between the two objects i and j. The core of the dataset is defined as:

$$\begin{aligned} Core(\mathbb{C}) \;\; = \;\; & \{a \in \mathbb{C} | \exists C_{ij}, \mu_{C_{ij}}(a) > 0, \\ & \forall f \in \{\mathbb{C} - a\} \mu_{C_{ij}}(f) = 0\} \end{aligned} \tag{37}$$

2) Fuzzy Discernibility Function: As with the crisp approach, the entries in the matrix can be used to construct the fuzzy discernibility function:

$$f_D(a_1^*, ..., a_m^*) = \wedge\{\vee C_{ij}^* | 1 \leq j < i \leq |\mathbb{U}|\} \tag{38}$$

where $C_{ij}^* = \{a_x^* | a_x \in C_{ij}\}$. The function returns values in $[0,1]$, which can be seen to be a measure of the extent to which the function is satisfied for a given assignment of truth values to variables. To discover reducts from the fuzzy discernibility function, the task is to find the minimal assignment of the value 1 to the variables such that the formula is maximally satisfied. By setting all variables to 1, the maximal value for the function can be obtained as this provides the most discernibility between objects.

Crisp discernibility matrices can be simplified by removing duplicate entries and clauses that are supersets of others. A similar degree of simplification can be achieved for fuzzy discernibility matrices. Duplicate clauses can be removed as a subset that satisfies one clause to a certain degree will always satisfy the other to the same degree.

3) Decision-relative Fuzzy Discernibility Matrix: As with the crisp discernibility matrix, for a decision system the decision feature must be taken into account for achieving reductions; only those clauses with different decision values are included in the crisp discernibility matrix. For the fuzzy version, this is encoded as:

$$\begin{aligned} f_D(a_1^*, ..., a_m^*) \;\; = \;\; & \{\wedge\{\{\vee C_{ij}^*\} \leftarrow q_{N(\mu_{R_q}(i,j))}\} | \\ & 1 \leq j < i \leq |\mathbb{U}|\} \end{aligned} \tag{39}$$

for decision feature q, where $\leftarrow$ denotes fuzzy implication. This construction allows the extent to which decision values differ to affect the overall satisfiability of the clause. If $\mu_{C_{ij}}(q)$ = 1 then this clause provides maximum discernibility (i.e. the two objects are maximally different according to the fuzzy similarity measure). When the decision is crisp and crisp equivalence is used, $\mu_{C_{ij}}(q)$ becomes 0 or 1.

4) Reduction: For the purposes of finding reducts, use of the fuzzy intersection of all clauses in the fuzzy discernibility function may not provide enough information for evaluating subsets. Here, it may be more informative to consider the individual satisfaction of each clause for a given set of features. The degree of satisfaction of a clause C_{ij} for a subset of features P is defined as:

$$SAT_P(C_{ij}) = \bigcup_{a \in P} \{\mu_{C_{ij}}(a)\} \tag{40}$$

Returning to the example, if the subset $P = \{a, c\}$ is chosen, the resulting degree of satisfaction of the clause is

$$SAT_P(C_{ij}) = \{0.4 \vee 0.2\} = 0.6$$

using the Łukasiewicz t-conorm, $\min(1, x + y)$.

For the decision-relative fuzzy indiscernibility matrix, the decision feature q must be taken into account also:

$$SAT_{P,q}(C_{ij}) = SAT_P(C_{ij}) \leftarrow \mu_{C_{ij}}(q) \tag{41}$$

For the example clause, if the corresponding decision values are crisp and are different, the degree of satisfaction of the clause is

$$\begin{aligned} SAT_{P,q}(C_{ij}) \;\; & = \;\; SAT_P(C_{ij}) \leftarrow 1 \\ & = \;\; 0.6 \leftarrow 1 \\ & = \;\; 0.6 \end{aligned}$$

For a subset P, the total satisfiability of all clauses can be calculated as

$$SAT(P) = \frac{\sum\limits_{i,j \in \mathbb{U}, i \neq j} SAT_{P,q}(C_{ij})}{\sum\limits_{i,j \in \mathbb{U}, i \neq j} SAT_{\mathbb{C},q}(C_{ij})} \tag{42}$$

where $\mathbb{C}$ is the full set of conditional attributes, and hence the denominator is a normalizing factor. If this value reaches 1 for a subset P, then the subset is a fuzzy-rough reduct. A proof of the monotonicity of the function $SAT(P)$ can be found in the appendix.

Many methods available from the literature for the purpose of finding reducts for crisp discernibility matrices are applicable here also. The Johnson Reducer [18] is extended and used herein to illustrate the concepts involved. This is a simple greedy heuristic algorithm that is often applied to discernibility functions to find a single reduct. Subsets of features found by this process have no guarantee of minimality, but are generally of a size close to the minimal.

The algorithm begins by setting the current reduct candidate, P, to the empty set. Then, each conditional feature appearing in the discernibility function is evaluated according to the heuristic measure used. For the standard Johnson algorithm, this is typically a count of the number of appearances a feature makes within clauses; features that appear more frequently are considered to be more significant. The feature with the highest heuristic value is added to the reduct candidate and all clauses in the discernibility function containing this feature are removed. As soon as all clauses have been removed, the algorithm terminates and returns the subset P. P is assured to be a fuzzy-rough reduct as all clauses contained within the discernibility function have been addressed. However, as with the other approaches, the subset may not necessarily have minimal cardinality.

The complexity of the algorithm is the same as that of FRFS in that $O((n^2+n)/2)$ calculations of the evaluation function ($SAT(P)$) are performed in the worst case. Additionally, this approach requires the construction of the fuzzy discernibility matrix, which has a complexity of $O(a * o^2)$ for a dataset containing a attributes and o objects.

5) Example: For the example dataset, the fuzzy discernibility matrix needs to be constructed based on the fuzzy discernibility given in equation (35) using the standard negator, and fuzzy similarity in equation (26). For objects 2 and 3, the resulting fuzzy clause is:

$$\{a_{0.301} \vee b_{1.0} \vee c_{0.964}\} \leftarrow q_{1.0}$$

where $\leftarrow$ denotes fuzzy implication. The fuzzy discernibility of objects 2 and 3 for attribute a is 0.301, indicating that the objects are partly discernible for this feature. The objects are fully discernible with respect to the decision feature, indicated by $q_{1.0}$. The full set of clauses is:

$$\begin{array}{lll}
C_{12}: & \{a_{0.0} \vee b_{1.0} \vee c_{1.0}\} & \leftarrow q_{1.0} \\
C_{13}: & \{a_{0.301} \vee b_{0.432} \vee c_{0.964}\} & \leftarrow q_{0.0} \\
C_{14}: & \{a_{1.0} \vee b_{0.0} \vee c_{1.0}\} & \leftarrow q_{1.0} \\
C_{15}: & \{a_{1.0} \vee b_{0.0} \vee c_{1.0}\} & \leftarrow q_{1.0} \\
C_{16}: & \{a_{1.0} \vee b_{1.0} \vee c_{1.0}\} & \leftarrow q_{0.0} \\
C_{23}: & \{a_{0.301} \vee b_{1.0} \vee c_{0.964}\} & \leftarrow q_{1.0} \\
C_{24}: & \{a_{1.0} \vee b_{1.0} \vee c_{0.482}\} & \leftarrow q_{0.0} \\
C_{25}: & \{a_{1.0} \vee b_{1.0} \vee c_{0.482}\} & \leftarrow q_{0.0} \\
C_{26}: & \{a_{1.0} \vee b_{0.863} \vee c_{0.482}\} & \leftarrow q_{1.0} \\
C_{34}: & \{a_{1.0} \vee b_{0.431} \vee c_{1.0}\} & \leftarrow q_{1.0} \\
C_{35}: & \{a_{1.0} \vee b_{0.431} \vee c_{1.0}\} & \leftarrow q_{1.0} \\
C_{36}: & \{a_{1.0} \vee b_{1.0} \vee c_{1.0}\} & \leftarrow q_{0.0} \\
C_{45}: & \{a_{0.301} \vee b_{0.0} \vee c_{0.0}\} & \leftarrow q_{0.0} \\
C_{46}: & \{a_{0.301} \vee b_{1.0} \vee c_{0.0}\} & \leftarrow q_{1.0} \\
C_{56}: & \{a_{0.0} \vee b_{1.0} \vee c_{0.0}\} & \leftarrow q_{1.0}
\end{array}$$

The feature selection algorithm then proceeds in the following way. Each individual feature is evaluated according to the measure defined in equation (42). For feature a, this is:

$$\begin{aligned}
SAT(\{a\}) &= \frac{\sum\limits_{i,j \in \mathbb{U}, i \neq j} SAT_{\{a\},q}(C_{ij})}{\sum\limits_{i,j \in \mathbb{U}, i \neq j} SAT_{\mathbb{C},q}(C_{ij})} \\
&= \frac{11.601}{15} \\
&= 0.773
\end{aligned}$$

Similarly for the remaining features:

$$SAT(\{b\}) = 0.782 \quad SAT(\{c\}) = 0.830$$

The feature that produces the largest increase in satisfiability is c. This feature is added to the reduct candidate, and the search continues:

$$SAT(\{a,c\}) = 0.887 \quad SAT(\{b,c\}) = 1.0$$

The subset $\{b,c\}$ is found to satisfy all clauses maximally, and the algorithm terminates. This subset is a fuzzy-rough reduct.

IV. Experimentation

This section presents the initial experimental evaluation of the selection methods for the task of pattern classification, over nine benchmark datasets from [4] and [13] with two classifiers.

A. Experimental Setup

FRFS uses a pre-categorization step which generates associated fuzzy sets for a dataset. For the new fuzzy-rough methods, the Łukasiewicz fuzzy connectives are used, with fuzzy similarity defined in (26). After feature selection, the datasets are reduced according to the discovered reducts. These reduced datasets are then classified using the relevant classifier. (Obviously, the feature selection step is not employed for the unreduced dataset.)

Two classifiers were employed for the purpose of evaluating the resulting subsets from the feature selection phase: JRip [6] and PART [34], [35]. JRip learns propositional rules by repeatedly growing rules and pruning them. During the growth phase, features are added greedily until a termination condition

TABLE III
REDUCT SIZE AND TIME TAKEN

Dataset	Objects	Features	Reduct size				Time taken (s)				
			FRFS	B-FRFS	L-FRFS	FDM	FRFS	B-FRFS	L-FRFS	FDM setup	FDM
Cleveland	297	14	11	9	9	9	24.11	8.78	3.32	8.75	1.93
Glass	214	10	9	9	10	9	1.61	3.30	1.53	4.28	0.60
Heart	270	14	11	8	8	8	11.84	3.61	2.17	7.31	1.46
Ionosphere	230	35	11	9	9	8	61.80	8.53	3.77	14.09	3.45
Olitos	120	26	10	6	6	6	11.20	1.29	0.72	3.61	0.46
Water 2	390	39	11	7	7	7	96.58	21.37	12.12	43.44	18.48
Water 3	390	39	12	7	7	7	158.73	27.36	13.44	44.43	16.95
Web	149	2557	24	20	21	18	5642.65	949.69	541.85	357.11	1425.58
Wine	178	14	10	6	6	6	1.42	1.69	0.97	4.16	0.44

is satisfied. Features are then pruned in the next phase subject to a pruning metric. Once the ruleset is generated, a further optimization is performed where classification rules are evaluated and deleted based on their performance on randomized data. PART generates rules by means of repeatedly creating partial decision trees from data. The algorithm adopts a divide-and-conquer strategy such that it removes instances covered by the current ruleset during processing. Essentially, a classification rule is created by building a pruned tree for the current set of instances; the leaf with the highest coverage is promoted to a rule.

B. Experimental Results

Table III compares the reduct size and runtime data for FRFS, fuzzy boundary region-based FS (B-FRFS), fuzzy lower approximation-based FS (L-FRFS) and fuzzy discernibility matrix-based FS (FDM). It can be seen that the new fuzzy-rough methods find smaller subsets than FRFS in general. The fuzzy boundary region-based method finds smaller or equally-sized subsets than the L-FRFS. This is to be expected, as B-FRFS includes fuzzy upper approximation information in addition to that of the fuzzy lower approximation. Of all the methods, the fuzzy discernibility matrix-based approach finds the smallest fuzzy-rough reducts. It is often seen in crisp rough set FS that discernibility matrix-based approaches find smaller subsets on average than those that rely solely on dependency degree information. This comes at the expense of setup time as can be seen in the table. Fuzzy clauses must be generated for every pair of objects in the dataset. The new fuzzy-rough methods are also quicker in computing reducts than FRFS, due mainly to the computation of the Cartesian product of fuzzy equivalence classes that FRFS must perform.

FRFS has been experimentally evaluated with other leading FS methods (such as Relief-F and entropy-based approaches [12], [13]) and has been shown to outperform these in terms of resulting classification performance. Hence, only comparisons to FRFS are given here. Table IV shows the average classification accuracy as a percentage obtained using 10-fold cross validation. The classification was initially performed on the unreduced dataset, followed by the reduced datasets which were obtained using the feature selection techniques. All techniques perform similarly, with classification accuracy improving or remaining the same for most datasets. FRFS performs equally well, however this is at the cost of extra features and extra time required to find reducts. The performance of the FDM method is generally slightly worse than the other methods. This can be attributed partly to the fact that the method produces smaller subsets for data reduction.

V. CONCLUSIONS

This paper has presented three new techniques for fuzzy-rough feature selection based on the use of fuzzy T-transitive similarity relations, that alleviate problems encountered with FRFS. The first development, based on fuzzy lower approximations, uses the similarity relations to construct approximations of decision concepts and evaluates these through a new measure of feature dependency. The second development employs the information in the fuzzy boundary region to guide the feature selection search process. When this is minimized, a fuzzy-rough reduct has been obtained. The third development extends the concept of the discernibility matrix to the fuzzy case, allowing features to belong to entries to a certain degree. An example FS algorithm is given to illustrate how reductions may be achieved. Note that no user-defined thresholds are required for any of the methods, although a choice must be made regarding fuzzy similarity relations and connectives.

Further work in this area will include a more in-depth experimental investigation of the proposed methods and the impact of the choice of relations and connectives. Additionally, the development of fuzzy discernibility matrices here allows the extension of many existing crisp techniques for the purposes of finding fuzzy-rough reducts. In particular, by reformulating the reduction task in a propositional satisfiability (SAT) framework, SAT solution techniques may be applied that should be able to discover such subsets, guaranteeing their minimality. The performance may also be improved through simplifying the fuzzy discernibility function further. This could be achieved by considering the properties of the fuzzy connectives and removing clauses that are redundant in the presence of others.

ACKNOWLEDGMENTS

This work is partly funded by the UK EPSRC grant GR/S98603/01. The authors are very grateful to Professor Colin Aitken and Mr Burkhard Schafer of the University of Edinburgh for their support.

TABLE IV
RESULTING CLASSIFICATION ACCURACIES (%)

Dataset	JRip					PART				
	Unred.	FRFS	B-FRFS	L-FRFS	FDM	Unred.	FRFS	B-FRFS	L-FRFS	FDM
Cleveland	52.19	54.55	54.55	54.55	54.55	50.17	52.19	53.20	53.20	53.20
Glass	71.50	69.63	65.89	71.50	65.89	67.76	68.22	70.56	70.56	67.76
Heart	77.41	78.89	78.52	78.52	78.52	73.33	78.52	76.30	76.30	76.30
Ionosphere	86.52	87.83	88.26	88.26	86.96	88.26	91.30	86.09	86.09	85.23
Olitos	70.83	70.83	71.67	64.17	63.33	57.50	62.50	67.50	58.33	64.17
Water 2	83.85	84.36	85.64	85.64	82.82	83.08	82.31	84.62	84.62	78.97
Water 3	82.82	82.82	79.74	81.28	80.00	83.33	80.51	80.26	79.23	79.74
Web	58.39	58.39	43.62	55.03	44.97	42.95	63.09	52.35	57.72	44.97
Wine	92.70	89.33	95.50	95.50	88.20	93.82	93.82	94.38	94.38	94.38

APPENDIX

Theorem 1: L-FRFS monotonicity. Suppose that $P \subseteq \mathbb{C}$, a is an arbitrary conditional feature that belongs to the dataset and Q is the set of decision features. Then $\gamma'_{P\cup\{a\}}(Q) \geq \gamma'_P(Q)$.

Proof: The fuzzy lower approximation of a concept X is

$$\mu_{\underline{R_{P\cup\{a\}}}X}(x) = \inf_{y\in\mathbb{U}} I(\mu_{R_{P\cup\{a\}}}(x,y), \mu_X(y))$$

From (23), it can be seen that

$$\mu_{R_{P\cup\{a\}}}(x,y) = \mu_{R_a}(x,y) \wedge \mu_{R_P}(x,y)$$

From the properties of t-norms, it can be seen that $\mu_{R_{P\cup\{a\}}}(x,y) \leq \mu_{R_P}(x,y)$. Thus, $I(\mu_{R_{P\cup\{a\}}}(x,y), \mu_X(y)) \geq I(\mu_{R_P}(x,y), \mu_X(y)), \forall x,y \in \mathbb{U}, X \in \mathbb{U}/Q$, and hence $\mu_{\underline{R_{P\cup\{a\}}}X}(x) \geq \mu_{\underline{R_P}X}(x)$. The fuzzy positive region of X is

$$\mu_{POS_{R_{P\cup\{a\}}}(Q)}(x) = \sup_{X\in\mathbb{U}/Q} \mu_{\underline{R_{P\cup\{a\}}}X}(x)$$

so $\mu_{POS_{R_{P\cup\{a\}}}(Q)}(x) \geq \mu_{POS_{R_P}(Q)}(x)$ and therefore $\gamma'_{P\cup\{a\}}(Q) \geq \gamma'_P(Q)$. ■

Theorem 2: B-FRFS monotonicity. Suppose that $P \subseteq \mathbb{C}$, a is an arbitrary conditional feature that belongs to the dataset and Q is the set of decision features. Then $\lambda_{P\cup\{a\}}(Q) \leq \lambda_P(Q)$.

Proof: The fuzzy boundary region of a concept X for an object x and set of features $P \cup \{a\}$ is defined as

$$\mu_{BND_{R_{P\cup\{a\}}}(X)}(x) = \mu_{\overline{R_{P\cup\{a\}}}X}(x) - \mu_{\underline{R_{P\cup\{a\}}}X}(x)$$

For the fuzzy upper approximation component of the fuzzy boundary region:

$$\mu_{\overline{R_{P\cup\{a\}}}X}(x) = \sup_{y\in\mathbb{U}} T(\mu_{R_{P\cup\{a\}}}(x,y), \mu_X(y))$$

It is known from Theorem 1 that $\mu_{R_{P\cup\{a\}}}(x,y) \leq \mu_{R_P}(x,y)$, so $\mu_{\overline{R_{P\cup\{a\}}}X}(x) \leq \mu_{\overline{R_P}X}(x)$. As $\mu_{\underline{R_{P\cup\{a\}}}X}(x) \geq \mu_{\underline{R_P}X}(x)$, then $\mu_{BND_{R_{P\cup\{a\}}}(X)}(x) \leq \mu_{BND_{R_P}(X)}(x)$. Thus, $U_{P\cup\{a\}}(Q) \leq U_P(Q)$ and therefore $\lambda_{P\cup\{a\}}(Q) \leq \lambda_P(Q)$. ■

Theorem 3: FDM monotonicity. Suppose that $P \subseteq \mathbb{C}$, a is an arbitrary conditional feature that belongs to the dataset and Q is the set of decision features. Then $SAT(P \cup \{a\}) \geq SAT(P)$.

Proof: For a clause C_{ij}, the degree of satisfaction for a given set of features $P \cup \{a\}$ is:

$$\begin{aligned} SAT_{P\cup\{a\}}(C_{ij}) &= \bigcup_{z\in P\cup\{a\}} \{\mu_{C_{ij}}(z)\} \\ &= SAT_P(C_{ij}) \cup \mu_{C_{ij}}(a) \end{aligned}$$

derived from the properties of the t-conorm. Thus, $SAT_{P\cup\{a\}}(C_{ij}) \geq SAT_P(C_{ij})$ for all clauses. Hence $SAT_{P\cup\{a\},q}(C_{ij}) \geq SAT_{P,q}(C_{ij})$. The overall degree of satisfaction for subset $P \cup \{a\}$ is

$$SAT(P \cup \{a\}) = \frac{\sum_{i,j\in\mathbb{U},i\neq j} SAT_{P\cup\{a\},q}(C_{ij})}{\sum_{i,j\in\mathbb{U},i\neq j} SAT_{\mathbb{C},q}(C_{ij})}$$

The denominator is a normalizing factor and can be ignored. As $SAT_{P\cup\{a\},q}(C_{ij}) \geq SAT_{P,q}(C_{ij})$ for all clauses, then $\sum_{i,j\in\mathbb{U},i\neq j} SAT_{P\cup\{a\},q}(C_{ij}) \geq \sum_{i,j\in\mathbb{U},i\neq j} SAT_{P,q}(C_{ij})$. Therefore, $SAT(P \cup \{a\}) \geq SAT(P)$. ■

Theorem 4: FDM reducts are fuzzy-rough reducts. Suppose that $P \subseteq \mathbb{C}$, a is an arbitrary conditional feature that belongs to the dataset and Q is the set of decision features. If P maximally satisfies the fuzzy discernibility function then P is a fuzzy-rough reduct.

Proof: The fuzzy positive region for a subset P is

$$\mu_{POS_{R_P}(Q)}(x) = \sup_{X\in\mathbb{U}/Q} \inf_{y\in\mathbb{U}} \{\mu_{R_P}(x,y) \rightarrow \mu_X(y)\}$$

The dependency function is maximized when each x belongs maximally to the fuzzy positive region. Hence,

$$\inf_{x\in\mathbb{U}} \sup_{X\in\mathbb{U}/Q} \inf_{y\in\mathbb{U}} \{\mu_{R_P}(x,y) \rightarrow \mu_X(y)\}$$

is maximized only when P is a fuzzy-rough reduct. This can be rewritten as the following:

$$\inf_{x,y\in\mathbb{U}} \{\mu_{R_P}(x,y) \rightarrow \mu_{R_q}(x,y)\}$$

when using a fuzzy similarity relation in the place of crisp decision concepts, as $\mu_{[x]_R} = \mu_R(x,y)$ [9]. Each $\mu_{R_P}(x,y)$ is constructed from the t-norm of its constituent relations:

$$\inf_{x,y\in\mathbb{U}} \{T_{a\in P}(\mu_{R_a}(x,y)) \rightarrow \mu_{R_q}(x,y)\}$$

This may be reformulated as

$$\inf_{x,y\in\mathbb{U}} \{S_{a\in P}(\mu_{R_a}(x,y) \rightarrow \mu_{R_q}(x,y))\} \tag{43}$$

Considering the fuzzy discernibility matrix approach, the fuzzy discernibility function is maximally satisfied when

$$\{\wedge\{\{\vee\ C^*_{xy}\} \leftarrow q_{N(\mu_{R_q}(x,y))}\} | 1 \leq y < x \leq |\mathbb{U}|\}$$

is maximized. This can be rewritten as:

$$T_{x,y\in\mathbb{U}}(S_{a\in P}(N(\mu_{R_a}(x,y))) \leftarrow N(\mu_{R_q}(x,y)))$$

because each clause C_{xy} is generated by considering the fuzzy similarity of values of each pair of objects x, y. Through the properties of the fuzzy connectives, this may be rewritten as:

$$T_{x,y\in\mathbb{U}}(S_{a\in P}(\mu_{R_a}(x,y) \rightarrow \mu_{R_q}(x,y))) \tag{44}$$

When this is maximized, (43) is maximized and so the subset P must be a fuzzy-rough reduct.

■

Theorem 5: Fuzzy-rough negative region is always empty for crisp decisions. Suppose that $P \subseteq \mathbb{C}$ and Q is the set of decision features. If Q is crisp, then the fuzzy-rough negative region is empty.

Proof: The fuzzy-rough negative region for subset P is

$$\mu_{NEG_{R_P}}(x) = N(\sup_{X\in\mathbb{U}/Q} \mu_{\overline{R_P}X}(x))$$

For the negative region to be empty, all object memberships must be zero. Hence

$$\forall x, \sup_{X\in\mathbb{U}/Q} \mu_{\overline{R_P}X}(x) = N(0) = 1$$

Expanding this gives

$$\forall x, \sup_{X\in\mathbb{U}/Q} \sup_{y\in\mathbb{U}} T(\mu_{R_P}(x,y), \mu_X(y)) = 1$$

For this to be maximized, there must be a suitable X and y such that

$$\forall x, \exists X, \exists y, T(\mu_{R_P}(x,y), \mu_X(y)) = 1$$

Setting $y = x$, the above holds as the decisions are crisp, so each x must belong fully to one decision X, $\mu_X(x) = 1$. Therefore, the fuzzy-rough negative region is always empty for crisp decisions. When the decisions are fuzzy and $\sup_{x\in X} \mu_X(x) < 1$ then the fuzzy-rough negative region will be non-empty.

■

REFERENCES

[1] "Rough Sets and Current Trends in Computing," *Proc. Third Int'l Conf.*, J.J. Alpigini, J.F. Peters, J. Skowronek, and N. Zhong, eds., 2002.

[2] R.B. Bhatt and M. Gopal, "On fuzzy-rough sets approach to feature selection," *Pattern Recognition Letters*, vol. 26, no. 7, pp. 965–975, 2005.

[3] R.B. Bhatt and M. Gopal, "On the compact computational domain of fuzzy-rough sets," *Pattern Recognition Letters*, vol. 26, no. 11, pp. 1632–1640, 2005.

[4] C.L. Blake and C.J. Merz. UCI Repository of machine learning databases. Irvine, University of California, 1998. http://www.ics.uci.edu/~mlearn/.

[5] A. Chouchoulas and Q. Shen, "Rough Set-Aided Keyword Reduction for Text Categorisation," *Applied Artificial Intelligence*, vol. 15, no. 9, pp. 843–873, 2001.

[6] W.W. Cohen, "Fast effective rule induction," In *Proceedings of the 12th International Conference on Machine Learning*, pp. 115–123, 1995.

[7] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, no. 3, 1997.

[8] M. De Cock, C. Cornelis, and E.E. Kerre, "Fuzzy Rough Sets: The Forgotten Step," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 1, pp. 121–130, 2007.

[9] D. Dubois and H. Prade, "Putting Rough Sets and Fuzzy Sets Together," *Intelligent Decision Support*, pp. 203–232, 1992.

[10] I. Düntsch and G. Gediga, *Rough Set Data Analysis: A Road to Non-Invasive Knowledge Discovery*. Bangor: Methodos, 2000.

[11] R. Jensen and Q. Shen, "Fuzzy-Rough Attribute Reduction with Application to Web Categorization," *Fuzzy Sets and Systems*, vol. 141, no. 3, pp. 469–485, 2004.

[12] R. Jensen and Q. Shen, "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1457–1471, 2004.

[13] R. Jensen and Q. Shen, "Fuzzy-Rough Sets Assisted Attribute Selection," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 1, pp. 73–89, 2007.

[14] J. Komorowski, Z. Pawlak, L. Polkowski and A. Skowron, "Rough Sets: A Tutorial," In [19], pp. 3-98. 1999.

[15] P. Langley, "Selection of Relevant Features in Machine Learning," *Proc. AAAI Fall Symp. Relevance*, pp. 1–5, 1994.

[16] P. Lingras and R. Jensen, "Survey of Rough and Fuzzy Hybridization," *Proceedings of the 16th International Conference on Fuzzy Systems* (FUZZ-IEEE'07), pp. 125–130, 2007.

[17] H. S. Nguyen and A. Skowron, "Boolean Reasoning for Feature Extraction Problems," ISMIS pp. 117-126, 1997.

[18] A. Øhrn, "Discernibility and Rough Sets in Medicine: Tools and Applications," Department of Computer and Information Science. Trondheim, Norway, Norwegian University of Science and Technology: 239. 1999.

[19] *Rough-Fuzzy Hybridization: A New Trend in Decision Making*, S.K. Pal and A. Skowron, eds. Springer Verlag, 1999.

[20] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishing, 1991.

[21] "Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems," *Studies in Fuzziness and Soft Computing*, L. Polkowski, T.Y. Lin, and S. Tsumoto, eds., vol. 56, Physica-Verlag, 2000.

[22] L. Polkowski, "Rough Sets: Mathematical Foundations," *Advances in Soft Computing*, Physica Verlag, 2002.

[23] A.M. Radzikowska and E.E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets and Systems*, vol. 126, no. 2, pp. 137–155, 2002.

[24] Q. Shen and A. Chouchoulas, "A Fuzzy-Rough Approach for Generating Classification Rules," *Pattern Recognition*, vol. 35, no. 11, pp. 341–354, 2002.

[25] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

[26] A. Skowron and C. Rauszer, "The discernibility matrices and functions in Information Systems," In: [27], pp. 331-362. 1992.

[27] *Intelligent Decision Support*, R. Slowinski, ed., Kluwer Academic Publishers, 1992.

[28] R.W. Swiniarski and A. Skowron, "Rough Set Methods in Feature Selection and Recognition," *Pattern Recognition Letters*, vol. 24, no. 6, pp. 833–849, 2003.

[29] H. Thiele, "Fuzzy Rough Sets versus Rough Fuzzy Sets - An Interpretation and a Comparative Study Using Concepts of Modal Logics," Technical Report no. CI-30/98, Univ. of Dortmund, 1998.

[30] G.C.Y. Tsang, D. Chen, E.C.C. Tsang, J.W.T. Lee, and D.S. Yeung, "On attributes reduction with fuzzy rough sets," *Proc. 2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 3, pp. 2775–2780, 2005.

[31] M. Wallace, Y. Avrithis and S. Kollias, "Computationally efficient supt transitive closure for sparse fuzzy binary relations," *Fuzzy Sets and Systems*, vol. 157, no. 3, pp. 341–372, 2006.

[32] J. Wang and J. Wang, "Reduction Algorithms Based on Discernibility Matrix: The Ordered Attributes Method," *J. Comput. Sci. & Technol.*, vol. 16, no. 6, pp. 489–504, 2001.

[33] X.Z. Wang, Y. Ha, and D. Chen, "On the reduction of fuzzy rough sets," *Proc. 2005 International Conference on Machine Learning and Cybernetics*, vol. 5, pp. 3174–3178, 2005.

[34] I.H. Witten and E. Frank, "Generating Accurate Rule Sets Without Global Optimization," In *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, 1998.

[35] I.H. Witten and E. Frank, *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2000.

[36] W.Z. Wu and W.X. Zhang, "Constructive and axiomatic approaches of fuzzy approximation operators," *Information Sciences*, vol. 159, no.3-4, pp. 233–254, 2004.

[37] W.Z. Wu, Y. Leung, and J.S. Mi, "On characterizations of (I,T)-fuzzy rough approximation operators," *Fuzzy Sets and Systems*, vol. 154, no. 1, pp. 76–102, 2005.

[38] Y.Y. Yao, "Combination of rough and fuzzy sets based on α-level sets," in: T.Y. Lin, N. Cereone (Eds.), *Rough Sets and Data Mining: Analysis of Imprecise Data*, Kluwer Academic Publishers, pp. 301–321, 1997.
[39] Y.Y. Yao, "A Comparative Study of Fuzzy Sets and Rough Sets," *Information Sciences*, vol. 109, pp. 21–47, 1998.
[40] Y.Y. Yao, Y. Zhao and J. Wang, "On reduct construction algorithms," *Proceedings of the First International Conference on Rough Sets and Knowledge Technology* (RSKT06), 297–304, 2006.
[41] Y.Y. Yao, "Decision-theoretic rough set models," *Proceedings of the International Conference on Rough Sets and Knowledge Technology* (RSKT07), LNAI 4481, pp. 1–12, 2007.
[42] D.S. Yeung, D. Chen, E.C.C. Tsang, J.W.T. Lee, and W. Xizhao, "On the Generalization of Fuzzy Rough Sets," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 3, pp. 343–361, 2005.
[43] L.A. Zadeh, "Fuzzy Sets," *Information and Control*, vol. 8, pp. 338–353, 1965.
[44] L.A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Information Sciences*, vol. 8, pp. 199–249, 301–357; vol. 9: 43–80, 1975.

Richard Jensen received the B.Sc. degree in computer science from Lancaster University, U.K., and the M.Sc. and Ph.D. degrees in artificial intelligence from the University of Edinburgh, U.K. He is a Lecturer with the Department of Computer Science at the University of Wales, Aberystwyth, working in the Advanced Reasoning Group. His research interests include rough and fuzzy set theory; pattern recognition; information retrieval; feature selection; and swarm intelligence. He has published over 25 peer-refereed articles in these areas.

Qiang Shen received the B.Sc. and M.Sc. degrees in communications and electronic engineering from the National University of Defence Technology, China, and the Ph.D. degree in knowledge-based systems from Heriot-Watt University, Edinburgh, U.K. He is a professor with the Department of Computer Science at the University of Wales, Aberystwyth, and an honorary fellow at the University of Edinburgh. His research interests include fuzzy and imprecise modeling, model-based inference, pattern recognition, and knowledge refinement and reuse. Dr Shen is an associate editor of the IEEE Transactions on Fuzzy Systems and of the IEEE Transactions on Systems, Man, and Cybernetics (Part B), and an editorial board member of the Fuzzy Sets and Systems Journal amongst others. He has published over 180 peer-refereed papers in academic journals and conferences on topics within artificial intelligence and related areas.

LIST OF FIGURES

List of Tables