A Noise-Tolerant Approach to Fuzzy-Rough Feature Selection

Chris Cornelis and Richard Jensen

Abstract— **In rough set based feature selection, the goal is to omit attributes (features) from decision systems such that objects in different decision classes can still be discerned. A popular way to evaluate attribute subsets with respect to this criterion is based on the notion of dependency degree. In the standard approach, attributes are expected to be qualitative; in the presence of quantitative attributes, the methodology can be generalized using fuzzy rough sets, to handle gradual (in)discernibility between attribute values more naturally. However, both the extended approach, as well as its crisp counterpart, exhibit a strong sensitivity to noise: a change in a single object may significantly influence the outcome of the reduction procedure. Therefore, in this paper, we consider a more flexible methodology based on the recently introduced Vaguely Quantified Rough Set (VQRS) model. The method can handle both crisp (discrete-valued) and fuzzy (real-valued) data, and encapsulates the existing noise-tolerant data reduction approach using Variable Precision Rough Sets (VPRS), as well as the traditional rough set model, as special cases.**

I. Introduction

Fuzzy sets [1] and rough sets [2] address two important characteristics of imperfect data and knowledge: the former model vague information by expressing that objects belong to a set or relation to a given degree, while the latter provide approximations of concepts in the presence of incomplete information. To merge these notions into a joint theory that combines their mutual strengths has been the object of a hybridisation movement that emerged in the early 1990's with the seminal research of Dubois and Prade [3] and has flourished ever since [4]. Recently, cross-disciplinary research has also been boosted by the adoption of computing paradigms like granular computing (see e.g. [5]), with its focus on clustering information entities into granules in terms of similarity or indiscernibility, and soft computing [6], which has stressed the role of fuzzy sets and rough sets as partners, rather than as adversaries, within a panoply of practical applications.

At the heart of the synergy between fuzzy sets and rough sets are the definitions of lower and upper approximations of a fuzzy set A under a fuzzy relation R (see e.g. [7] for a fairly general version of these definitions). In this framework, R assesses objects' indiscernibility, such that objects are categorized into classes, or granules, with soft boundaries based on their similarity to one another. On the other hand, the fuzzy set A models a vague concept, i.e., such that objects can meet its characteristics to varying degrees.

C. Cornelis is with the Department of Applied Mathematics and Computer Science, Ghent University, Gent, Belgium (email: Chris.Cornelis@UGent.be)

R. Jensen is with the Department of Computer Science, Aberystwyth University, UK (email: rkj@aber.ac.uk)

The work of C. Cornelis was supported by the Research Foundation–Flanders.

Recently, it was noted in [8] that by focusing on conservative extensions of its contributing ingredients, fuzzy rough set theory inherits not only their strengths, but also some of their weaknesses. In particular, although they allow for gradual membership, the classical branch of fuzzy rough sets are still abrupt in a sense that adding or omitting a single element may drastically alter the outcome of the approximations. Therefore, the authors proposed vaguely quantified rough sets (VQRS), in which an object y belongs to the lower approximation of a set A to the extent that most objects related to y are in A, and to the upper approximation to the extent that some objects related to y are in A. The discerning feature of the VQRS approach is the introduction of vague quantifiers like 'some' or 'most' into the approximations; it extends Ziarko's noise-tolerant model of variable precision rough sets (VPRS, [9]), which uses crisp thresholds $0 \leq l < u \leq 1$ to add an element y to the lower approximation of a set A if at least $100 * u$ percent of the elements related to y are in A, and to its upper approximation if more than $100 * l$ percent of the elements related to y are in A.

In this paper, we explore the potential of the VQRS model for feature selection [10], [11] in decision systems, i.e., the problem of selecting those input features (attributes) that are most predictive of the outcome (decision) of the system. Rough set analysis [12] is very well-suited for this problem because it can achieve semantics-preserving data dimensionality reduction without the need for additional parameters other than the supplied data itself. The original framework requires that data be qualitative (discrete-valued, nominal or crisp); this means that quantitative (real-valued, continuous or fuzzy) data need to be preprocessed, either by replacing the numerical attribute values by interval codes (discretisation, see e.g. [13], [14]), or by considering a notion of approximate equality, or graded indiscernibility, between objects, leading to fuzzy-rough feature selection (FRFS) methods (see e.g. [15], [16]).

In either case, and in fact regardless of whether qualitative or quantitative data are used, noise is an important factor degrading the performance of reduction: a single misclassified object prevents rough set analysis from making any conclusive statements about all other objects it is related to. To reduce the impact of noise, the original rough set approach has been adapted by using VPRS approximations (see e.g. [17]), such that problematic elements are not taken into account as long as their relative proportion remains below a certain threshold. In this paper, we go one step further by relaxing this crisp threshold into a smoother region of tolerance towards classification errors. As an added benefit, our approach can be integrated seamlessly with FRFS approaches, providing a general model that encapsulates all

978-1-4244-1819-0/08/$25.00

the above-mentioned approaches as specific cases.

The remainder of this paper is structured as follows: in Section II, we review the fuzzy-rough hybridisation process by briefly recalling its ingredients (fuzzy sets and rough sets) as well as its resulting end products (fuzzy rough sets vs. vaguely quantified rough sets). Section III focuses on feature selection: after recalling the classical rough set based procedure for qualitative data reduction (Section III-A), as well as its fuzzy-rough extension to quantitative data, and the associated notion of fuzzy decision reducts [18] (section III-B), we outline the VQRS-based approach in Section III-C. We also investigate its theoretical characteristics; as with the VPRS approach, some basic properties taken for granted in the traditional case do not extend to the noise-tolerant setting, and practical implementations need to be aware of this. Initial experimental results that demonstrate the potential of the approach are presented in Section IV. Finally, Section V concludes the paper and outlines some ideas for future work.

II. Fuzzy-Rough Hybridisation

A. Fuzzy Sets

Recall that a fuzzy set in X is an $X \to [0,1]$ mapping, while a fuzzy relation in X is a fuzzy set in $X \times X$. For all y in X, the R-foreset of y is the fuzzy set Ry defined by $Ry(x) = R(x,y)$ for all x in X. If R is reflexive and symmetric, i.e., $R(x,x) = 1$ and $R(x,y) = R(y,x)$ hold for all x and y in X, then R is called a fuzzy tolerance relation. For fuzzy sets A and B in X, $A \subseteq B \iff (\forall x \in X)(A(x) \leq B(x))$. The intersection $A \cap B$ and union $A \cup B$ of A and B are defined in this paper by, for x in X,

$$(A \cap B)(x) = \min(A(x), B(x)) \quad (1)$$

$$(A \cup B)(x) = \max(A(x), B(x)) \quad (2)$$

If X is finite, the cardinality of A equals

$$|A| = \sum_{x \in X} A(x) \quad (3)$$

Fuzzy logic connectives play an important role in the hybridisation process. We therefore recall some important definitions. A triangular norm (t-norm for short) $\mathcal{T}$ is any increasing, commutative and associative $[0,1]^2 \to [0,1]$ mapping satisfying $\mathcal{T}(1,x) = x$, for all x in $[0,1]$. Common examples of t-norms include the minimum, the product and $\mathcal{T}_L$ defined by $\mathcal{T}_L(x,y) = \max(0, x+y-1)$ for x, y in [0,1]. An implicator is any $[0,1]^2 \to [0,1]$-mapping $\mathcal{I}$ that is decreasing in its first, and increasing in its second component, and that satisfies $\mathcal{I}(0,0) = 1$ and $\mathcal{I}(1,x) = x$, for all x in $[0,1]$. In this paper, we consider $\mathcal{I}_L$, defined by, for x, y in [0,1],

$$\mathcal{I}_L(x,y) = \min(1, 1-x+y) \quad (4)$$

It satisfies the following property, called confinement principle (see e.g. [19]), for x and y in $[0,1]$,

$$x \leq y \Leftrightarrow \mathcal{I}(x,y) = 1 \quad (5)$$

B. Rough Sets (RS)

Rough set theory makes statements about the membership of an object y of X to the concept of which A is a set of examples, based on the indiscernibility between y and the elements of A. Usually, indiscernibility is described by means of an equivalence relation R in X; in this case, (X,R) is called a standard, or Pawlak, approximation space. In a Pawlak approximation space (X,R), an element y of X belongs to the lower approximation $R{\downarrow}A$ of A if the equivalence class Ry of y is included in A. On the other hand, y belongs to the upper approximation $R{\uparrow}A$ of A if its equivalence class has a non-empty intersection with A:

$$y \in R{\downarrow}A \quad \text{iff} \quad Ry \subseteq A \quad (6)$$

$$y \in R{\uparrow}A \quad \text{iff} \quad Ry \cap A \neq \emptyset \quad (7)$$

In other words,

$$y \in R{\downarrow}A \quad \text{iff} \quad (\forall x \in X)((x,y) \in R \Rightarrow x \in A) \quad (8)$$

$$y \in R{\uparrow}A \quad \text{iff} \quad (\exists x \in X)((x,y) \in R \wedge x \in A) \quad (9)$$

C. Fuzzy Rough Sets (FRS)

Research on hybridising fuzzy sets and rough sets has focused mainly on fuzzifying the definitions of lower and upper approximation. Typically, it is assumed that R is at least a fuzzy tolerance relation.

For the lower and upper approximation of a fuzzy set A in X by means of R, we adopt the definitions proposed by Radzikowska and Kerre in [7]: given an implicator $\mathcal{I}$ and a t-norm $\mathcal{T}$, Formulas (8) and (9) are paraphrased to define $R{\downarrow_{\mathcal{I}}}A$ and $R{\uparrow_{\mathcal{T}}}A$ in X by

$$(R{\downarrow_{\mathcal{I}}}A)(y) = \inf_{x \in X} \mathcal{I}(R(x,y), A(x)) \quad (10)$$

$$(R{\uparrow_{\mathcal{T}}}A)(y) = \sup_{x \in X} \mathcal{T}(R(x,y), A(x)) \quad (11)$$

for all y in X.

D. Vaguely Quantified Rough Sets (VQRS)

Formulas (10) and (11) have been conceived with the purpose of conserving the traditional lower and upper approximations in mind. Indeed, when A and R are both crisp, it can be verified that (8) and (9) are recovered. Note in particular how the $\inf$ and $\sup$ operations play the same role as the $\forall$ and $\exists$ quantifiers, and how a change in a single element can thus have a large impact on (10) and (11). This makes fuzzy rough sets equally susceptible to noisy data — which is difficult to rule out in real-life applications— as their crisp counterparts.

To make up for this shortcoming, Cornelis et al. [8] proposed to soften the universal and existential quantifier by means of vague quantifiers like *most* and *some*. Mathematically, they modeled such vague quantifiers in terms of Zadeh's [20] notion of a regularly increasing fuzzy quantifier Q: an increasing $[0,1] \to [0,1]$ mapping that satisfies the boundary conditions $Q(0) = 0$ and $Q(1) = 1$.

Examples of fuzzy quantifiers can be generated by means of the following parametrized formula, for $0 \leq \alpha < \beta \leq 1$, and x in $[0,1]$,

$$Q_{(\alpha,\beta)}(x) = \begin{cases} 0, & x \leq \alpha \\ \frac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \alpha \leq x \leq \frac{\alpha+\beta}{2} \\ 1 - \frac{2(x-\beta)^2}{(\beta-\alpha)^2}, & \frac{\alpha+\beta}{2} \leq x \leq \beta \\ 1, & \beta \leq x \end{cases} \quad (12)$$

For instance, $Q_{(0.1,0.6)}$ and $Q_{(0.2,1)}$ might be used respectively to reflect the vague quantifiers *some* and *most* from natural language.

Once a couple (Q_l, Q_u) of fuzzy quantifiers is fixed, the Q_l-upper and Q_u-lower approximation of a fuzzy set A under a fuzzy relation R are defined by

$$(R\uparrow_{Q_l} A)(y) = Q_l\left(\frac{|Ry \cap A|}{|Ry|}\right) \quad (13)$$

$$(R\downarrow_{Q_u} A)(y) = Q_u\left(\frac{|Ry \cap A|}{|Ry|}\right) \quad (14)$$

for all y in X. In other words, an element y belongs to the lower approximation of A if most of the elements related to y are included in A. Likewise, an element belongs to the upper approximation of A if some of the elements related to y are included in A. Remark that when A and R are a crisp set and a crisp equivalence relation, respectively, the approximations may still be non-crisp. In this case, note also that when

$$Q_{>x_l}(x) = \begin{cases} 0, & x \leq x_l \\ 1, & x > x_l \end{cases} \qquad Q_{\geq x_u}(x) = \begin{cases} 0, & x < x_u \\ 1, & x \geq x_u \end{cases}$$

with $0 \leq x_l < x_u \leq 1$ are used as quantifiers, we recover Ziarko's variable precision rough set (VPRS) model [9], [21], and moreover when we use

$$Q_{\exists}(x) = \begin{cases} 0, & x = 0 \\ 1, & x > 0 \end{cases} \qquad Q_{\forall}(x) = \begin{cases} 0, & x < 1 \\ 1, & x = 1 \end{cases}$$

we obtain Pawlak's standard rough set model as a particular case of the VQRS approach.

As such, the VQRS model puts dealing with noisy data into an interesting new perspective: it inherits both the flexibility of VPRSs for dealing with classification errors (by relaxing the membership conditions for the lower approximation, and tightening those for the upper approximation) and that of fuzzy sets for expressing partial constraint satisfaction (by distinguishing different levels of membership to the upper/lower approximation).

III. Feature Selection

In the following, we assume that $(X, \mathcal{A} \cup \{d\})$ is a decision system, i.e., $X = \{x_1, \ldots, x_n\}$ and $\mathcal{A} = \{a_1, \ldots, a_m\}$ are finite, non-empty sets of objects and conditional attributes, respectively, and d is a designated attribute outside $\mathcal{A}$ called decision or class attribute. Each a in $\mathcal{A} \cup \{d\}$ corresponds to an $X \to V_a$ mapping, in which V_a is the value set of a over X. In general, value sets of all attributes can be infinite, but in this paper we assume that $V_d = \{v_1, \ldots, v_p\}$ $(p \geq 2)$; in this way, X is partitioned into p decision classes X_k $(k = 1, \ldots, p)$.

A. RS-Based Feature Selection

Central to rough set based attribute reduction is the concept of indiscernibility. For every subset B of $\mathcal{A} \cup \{d\}$, the B-indiscernibility relation R_B is defined as

$$R_B = \{(x,y) \in X^2 \text{ and } (\forall a \in B)(a(x) = a(y))\} \quad (15)$$

Clearly, each R_B is an equivalence relation. When $B \subseteq \mathcal{A}$, its equivalence classes can be used to approximate concepts, i.e., subsets A of X, by means of $R_B\downarrow A$ and $R_B\uparrow A$.

In practice, the concepts are usually equivalence classes of the decision attribute. Given $B \subseteq \mathcal{A}$, the B-positive region POS_B contains those objects for which the values of B allow to predict the decision class unequivocally:

$$POS_B = \bigcup_{k=1}^{p} R_B\downarrow X_k \quad (16)$$

The predictive ability w.r.t. d of the attributes in B is then measured by the following value (degree of dependency of d on B):

$$\gamma_B = \frac{|POS_B|}{|X|} \quad (17)$$

$(X, \mathcal{A} \cup \{d\})$ is called consistent if $\gamma_{\mathcal{A}} = 1$, i.e., if all objects are discernible when the entire conditional attribute set is taken into account. A subset B of $\mathcal{A}$ is called a decision reduct if it satisfies $POS_B = POS_{\mathcal{A}}$, i.e., B preserves the decision making power of $\mathcal{A}$, and if it cannot be further reduced, i.e., there exists no proper subset B' of B such that $POS_{B'} = POS_{\mathcal{A}}$. If the latter constraint is lifted, i.e., B is not necessarily minimal, we call B a decision superreduct.

Decision (super)reducts can be used to synthesize minimal decision rules: the rules result from overlaying the reducts over the original decision system and reading off the values. Unfortunately, computing all decision reducts is an NP-complete problem. In many cases, however, it suffices to generate a single (super)reduct of a decision system, a problem for which several heuristic algorithms have been devised.

For instance, the QuickReduct algorithm [16], [22], shown in Algorithm 1, starts off with $B = \emptyset$, and computes $\gamma_{B \cup \{a\}}$ for each attribute a in $\mathcal{A}$; the attribute for which this value is highest (or one of them in case there are several) is selected and added to B. Then, the same process is repeated for the remaining attributes, until $\gamma_B = \gamma_{\mathcal{A}}$. ReverseReduct [15], shown in Algorithm 2, proceeds in a dual fashion, starting with $B = \mathcal{A}$, and progressively eliminating attributes from B as long as $\gamma_B = \gamma_{\mathcal{A}}$.

Note that, by construction, when ReverseReduct finishes, the set B is guaranteed to equal a true reduct of $(X, \mathcal{A} \cup \{d\})$. This does not hold, in general, for QuickReduct, which may produce a superreduct B, i.e., while $\gamma_B = \gamma_{\mathcal{A}}$, there may be proper subsets B' of B also satisfying this property. In practice, when $\mathcal{A}$ is very large (e.g., contains hundreds of attributes), ReverseReduct may be computationally expensive, or even infeasible, because the construction of B-indiscernibility relations for large subsets B of $\mathcal{A}$ is very time-consuming.

Algorithm 1: The QUICKREDUCT Algorithm

(1) $B \leftarrow \{\}$
(2) **repeat**
(3) $T \leftarrow B$
(4) **foreach** $a \in (\mathcal{A} \setminus B)$
(5) **if** $\gamma_{B\cup\{a\}} > \gamma_T$
(6) $T \leftarrow B \cup \{a\}$
(7) $B \leftarrow T$
(8) **until** $\gamma_B = \gamma_{\mathcal{A}}$
(9) **return** B

Algorithm 2: The REVERSEREDUCT algorithm

(1) $B \leftarrow \mathcal{A}$
(2) **repeat**
(3) $T \leftarrow \emptyset$
(4) **foreach** $a \in B$
(5) **if** $\gamma_{B\setminus\{a\}} = \gamma_{\mathcal{A}}$
(6) $T \leftarrow B \setminus \{a\}$
(7) $B \leftarrow T$
(8) **until** $T = \emptyset$
(9) **return** B

B. FRS-Based Feature Selection

The RS-based feature selection approach requires that the value sets of all attributes in a decision system be finite. In order to cope with objects described by quantitative measurements, it is possible to use discretisation, yet often it is more natural, and more effective, to consider a gradual notion of discernibility rather than an absolute one [18].

There are several ways of constructing fuzzy (tolerance) relations (see e.g. [23]) that express the extent to which two objects are indiscernible. In this paper, given a quantitative attribute a, we compute the approximate equality between two objects w.r.t. a, by the relation R_a [16], defined by, for x and y in X:

$$R_a(x,y) = \max\left(0, 1 + \frac{\min(a(y)-a(x), a(x)-a(y))}{\sigma_a}\right) \quad (18)$$

in which σ_a^2 represents the variance of attribute a.

Assuming that for a qualitative attribute a, the classical way of discerning objects is used, i.e., $R_a(x,y) = 1$ if $a(x) = a(y)$ and $R_a(x,y) = 0$ otherwise, for any subset B of $\mathcal{A}$, the fuzzy B-indiscernibility relation R_B is defined by conjunctively combining the individual fuzzy relations R_a ($a \in B$) with a t-norm $\mathcal{T}$.

It can easily be seen that R_B is a fuzzy tolerance relation, and also that if only qualititative attributes (possibly stemming from discretisation) are used, then the traditional concept of B-indiscernibility relation is recovered.

Using fuzzy B-indiscernibility relations, the fuzzy B-positive region [16], [18] is defined by, for y in X,

$$POS_B(y) = \left(\bigcup_{k=1}^{p} R_B{\downarrow}X_k\right)(y) \quad (19)$$

Hence, the fuzzy B-positive region is a fuzzy set in X, to which an object y belongs to the extent that its R_B-foreset is included into at least one of the decision classes. As shown in [18], Formula (19) can be simplified to

$$POS_B(y) = (R_B{\downarrow}X_{k^*})(y) \quad (20)$$

such that $d(y) = v_{k^*}$. In other words, to determine the membership of y to the fuzzy B-positive region, only the decision class y belongs to needs to be inspected.

In [18], a general notion of fuzzy decision reduct based on an increasing $[0,1]$-valued measure was introduced. In this paper, we consider a particular instantiation of this definition based on a normalized[1] extension of the degree of dependency:

$$\gamma_B = \frac{|POS_B|}{|POS_{\mathcal{A}}|} \quad (21)$$

B is called a fuzzy decision superreduct to degree α if $\gamma_B \geq \alpha$, and a fuzzy decision reduct to degree α if moreover for all $B' \subset B$, $\gamma_{B'} < \alpha$.

In order to produce a single fuzzy decision (super)reduct to a preset degree α ($\alpha \in]0,1]$), we can use modified versions of QUICKREDUCT and REVERSEREDUCT, shown in Algorithms 3 and 4. Again, by construction, QUICKREDUCT produces guaranteed fuzzy decision superreducts, while REVERSEREDUCT obtains fuzzy decision reducts.

Algorithm 3: The fuzzy-rough QUICKREDUCT algorithm ($\alpha \in]0,1]$)

(1) $B \leftarrow \{\}$
(2) **repeat**
(3) $T \leftarrow B$
(4) **foreach** $a \in (\mathcal{A} \setminus B)$
(5) **if** $\gamma_{B\cup\{a\}} > \gamma_T$
(6) $T \leftarrow B \cup \{a\}$
(7) $B \leftarrow T$
(8) **until** $\gamma_B \geq \alpha$
(9) **return** B

Algorithm 4: The REVERSEREDUCT algorithm ($\alpha \in]0,1]$)

(1) $B \leftarrow \mathcal{A}$
(2) **repeat**
(3) $T \leftarrow \emptyset$; $\beta \leftarrow \alpha$
(4) **foreach** $a \in B$
(5) **if** $\gamma_{B\setminus\{a\}} \geq \beta$
(6) $T \leftarrow B \setminus \{a\}$
(7) $\beta \leftarrow \gamma_{B\setminus\{a\}}$
(8) $B \leftarrow T$
(9) **until** $T = \emptyset$
(10) **return** B

[1]Normalization is required in order that the measure yield a value of 1 for the whole attribute set; in this way, the notion of fuzzy reduct to degree α is meaningful regardless of the consistency of the decision system. In this paper, we assume $POS_{\mathcal{A}} \neq \emptyset$.

C. VQRS-Based Feature Selection

We start this section by presenting a few simple examples to illustrate the negative effects of noise on feature selection. We explain how the existing VPRS model tackles these defects, and then demonstrate how using VQRS lower approximation can extend the limited facilities of the VPRS approach to provide a finer-grained and more flexible noise handling mechanism.

Example 1: Consider the decision system D_1 in Table I. It is discrete-valued and has two decision reducts, viz. $\{a_1\}$ and $\{a_2, a_3\}$. When we corrupt the system by changing the decision of x_5 to 1 (resulting in D_2 in Table I), the decision reducts are $\{a_1, a_2\}$ and $\{a_2, a_3\}$. In other words, the noise has increased the average reduct size, and there is no longer a reduct of length 1 in the corrupted decision system. Moreover, $POS_{\{a_1\}} = \{x_2, x_3, x_4\}$, so $\gamma_{\{a_1\}} = 3/7$, a very sharp drop in the dependency degree considering that only one element out of seven has been affected.

A certain tolerance to noise may be introduced by using Ziarko's VPRS model; for instance, if we use $x_u = 0.75$ as a threshold, and replace the lower approximation in definition (16) of the positive region by the corresponding VPRS lower approximation $\downarrow_{Q_{\geq 0.75}}$, then the $\{a_1\}$-positive region contains every object in D_2. For instance, x_1 belongs to this positive region since

$$\frac{|R_{\{a_1\}}x_1 \cap X_0|}{|R_{\{a_1\}}x_1|} = \frac{|\{x_1, x_6, x_7\}|}{|\{x_1, x_5, x_6, x_7\}|} = \frac{3}{4} \qquad (22)$$

Hence, with this definition, $\gamma_{\{a_1\}} = 1$.

TABLE I
A) DECISION SYSTEM D_1 B) DECISION SYSTEM D_2.

	a_1	a_2	a_3	d
x_1	0	1	0	0
x_2	2	1	1	1
x_3	1	2	1	0
x_4	2	0	1	1
x_5	0	2	0	0
x_6	0	0	0	0
x_7	0	1	0	0

	a_1	a_2	a_3	d
x_1	0	1	0	0
x_2	2	1	1	1
x_3	1	2	1	0
x_4	2	0	1	1
x_5	0	2	0	1
x_6	0	0	0	0
x_7	0	1	0	0

This example confirms the use of the VPRS model, but it is clear that the choice of the threshold is crucial — if a slightly higher value of l is chosen, say $l = 0.8$, the initial problems reappear, and if l is chosen too low, say $l = 0.65$, $\{a_3\}$ is returned as a reduct by the reduction procedure as well. Moreover, as the following example shows, it also makes a difference which particular object is affected by noise.

Example 2: Consider the decision system D_3 in Table II, which is the same as D_1 but with x_4 changed instead of x_5. Using classical lower approximation, the dependency degree for $\{a_1\}$ is now $\gamma_{\{a_1\}} = 5/7$, a higher value than in Ex. 1. However, this value does not increase when using the VPRS lower approximation $\downarrow_{Q_{\geq 0.75}}$, since e.g.

$$\frac{|R_{\{a_1\}}x_2 \cap X_1|}{|R_{\{a_1\}}x_2|} = \frac{|\{x_2\}|}{|\{x_1, x_2\}|} = \frac{1}{2} < 0.75 \qquad (23)$$

TABLE II
DECISION SYSTEM D_3.

	a_1	a_2	a_3	d
x_1	0	1	0	0
x_2	2	1	1	1
x_3	1	2	1	0
x_4	2	0	1	0
x_5	0	2	0	0
x_6	0	0	0	0
x_7	0	1	0	0

The examples indicate that noise negatively impacts decision systems, since longer reducts also mean less general, weaker rules, and that, as a noise-handling mechanism, the VPRS approach is useful but rather opaque when it comes to choosing the right threshold. It is also fairly coarse-grained, classifying objects either as belonging to the positive region or not. These observations motivate the need for a smoother approach, based on the following VQRS-based definition of positive region:

$$POS_B^{Q_u}(y) = \left(\bigcup_{k=1}^{p} R_B \downarrow_{Q_u} X_k \right)(y) \qquad (24)$$

The VQRS degree of dependency of d on B, $\gamma_B^{Q_u}$, can be defined analogously as in Formula (21), but some precautions apply, see further on. Similarly as in FRS-based feature selection, it is possible to look for fuzzy decision reducts to a certain degree (regardless of whether the data is qualitative, quantitative or mixed).

Example 3: If $Q_u = Q_{(0.25,0.75)}$ is used in Formula (24), the $\{a_1\}$-positive region contains all objects of D_2, so $\gamma_{\{a_1\}}^{Q_u} = 1$. Also, $\gamma_{\{a_2\}}^{Q_u} \approx 0.69$ and $\gamma_{\{a_3\}}^{Q_u} \approx 0.97$. For D_3, it can be seen that e.g. $POS_{\{a_1\}}^{Q_u}(x_2) = Q_u(1/2) = 1/2$, and that $\gamma_{\{a_1\}}^{Q_u} = 6/7$. Likewise, it can be verified that $\gamma_{\{a_2\}}^{Q_u} = \gamma_{\{a_3\}}^{Q_u} \approx 0.98$. This indicates that, for this data, $\{a_2\}$ and $\{a_3\}$ are better candidates for data reduction than $\{a_1\}$.

It is important to note that, unlike in the FRS-based model of the previous section, a simplification like the one in Formula (20) does not apply automatically. Indeed, if an object is misclassified (like x_5 in D_2), it is likely to belong to a larger extent to the lower approximation of another decision class (in this case, X_0) than to that of its own class (i.e., X_1).

Still, there may be reasons to prefer such a simplification over the general definition (24). From a pragmatic perspective, the reduction in computational cost is significant (computing lower approximations for one decision class vs. for all of them). Also, thinking about rule induction, it would be undesirable to induce rules from misclassified objects (e.g., "if $a_1 = 0$ then $d = 1$" in the case of x_5 and D_2). This can be prevented if such an object is excluded from the positive region, which effectively happens when we replace Formula (24) by

$$POS_B'^{Q_u}(y) = (R_B \downarrow_{Q_u} X_{k^*})(y) \qquad (25)$$

with k^* such that $d(y) = v_{k^*}$. The corresponding degree of dependency is denoted $\gamma'^{Q_u}_B$.

Example 4: Consider again the data of Ex. 1. If we use Formula (25) to compute the positive region, $POS'^{Q_u}_{\{a_1\}}(x_1) = Q_u(3/4) = 1$, but $POS'^{Q_u}_{\{a_1\}}(x_5) = Q_u(1/4) = 0$, and $\gamma'^{Q_u}_{\{a_1\}} \approx 0.86$. Similarly, $\gamma'^{Q_u}_{\{a_2\}} \approx 0.56$ and $\gamma'^{Q_u}_{\{a_3\}} \approx 0.71$.

Clearly, $POS'^{Q_u}_B \subseteq POS^{Q_u}_B$ always holds. Unfortunately, neither POS^{Q_u} or POS'^{Q_u} itself is monotonic, that is, from $B_1 \subseteq B_2$ does not always follow that $POS^{Q_u}_{B_1} \subseteq POS^{Q_u}_{B_2}$, nor $POS'^{Q_u}_{B_1} \subseteq POS'^{Q_u}_{B_2}$. As a consequence, γ^{Q_u} and γ'^{Q_u} are not monotonic, either.

Example 5: Consider again the data of Ex. 3. It can be verified that $\gamma_{\{a_2\}} \approx 0.98 > \gamma_{\{a_1,a_2\}} \approx 0.86$.

Non-monotonicity of the dependency degree, which occurs also in the VPRS approach, generates a number of complications that are both of theoretical and practical concern. First, it can occur that $POS^{Q_u}_B \not\subseteq POS^{Q_u}_{\mathcal{A}}$. For such a subset B, computing $\gamma^{Q_u}_B$ as in Formula (21) results in a dependency degree that is strictly greater than 1. Therefore, a safer way of defining $\gamma^{Q_u}_B$ is given by

$$\gamma^{Q_u}_B = \min\left(1, \frac{|POS^{Q_u}_B|}{|POS^{Q_u}_{\mathcal{A}}|}\right) \tag{26}$$

Similar observations can be made for $\gamma'^{Q_u}_B$. Note that the above problems do not occur when the decision system is consistent.

Non-monotonicity also restricts the effectiveness of heuristic algorithms like (fuzzy-rough) QUICKREDUCT and REVERSEREDUCT, in a sense that neither of them is guaranteed to produce true (fuzzy) decision reducts. While from a theoretical point of view this is a fairly heavy price to pay, in practice the algorithms can still be used to produce sufficiently good attribute subsets.

Finally, the following proposition reveals an interesting relationship between the FRS- and the VQRS-based approach in case the second parameter of the fuzzy quantifier $Q_{(\alpha,\beta)}$ is equal to 1.

Proposition 1: Assume $(X, \mathcal{A} \cup \{d\})$ is a consistent decision system, $B \subseteq \mathcal{A}$, $Q_u = Q_{(\alpha,1)}$, and $\mathcal{I}$ is an implicator that satisfies the confinement principle (5). Then $\gamma^{Q_u}_B = 1$ if and only if $\gamma_B = 1$.

The proposition can be used to force the VQRS reduction process to generate fuzzy decision (super)reducts to degree 1 in the sense of Section III-B. In this way, if QUICKREDUCT is used, the noise-handling facilities of the VQRS approach are applied only in the intermediary stages, in which, depending on the value of α in $Q_{(\alpha,1)}$, a more flexible attribute selection criterion is used.

IV. EXPERIMENTAL RESULTS

In this section, we perform a number of preliminary experiments to analyse the performance of the VQRS feature selection approach for the task of classification, and to investigate the role of certain parameters and options. In particular, we compare the alternatives (24) and (25) for defining the VQRS positive region, and we consider two choices of fuzzy quantifiers, viz. $Q_u = Q_{(0,0.8)}$ and $Q_u = Q_{(0.2,1)}$. We compare them to the VPRS approach, using $x_u = 0.9$ as a threshold, and to the traditional RS- and FRS-based reduction approaches. Due to space restrictions, we only apply (fuzzy-rough) QUICKREDUCT for generating attribute subsets, and we always put the treshold α equal to 1, i.e., the algorithm finishes when a subset is found with a dependency degree (γ, γ^{Q_u} or γ'^{Q_u}, depending on the approach) of 1.

To back up the claim that our approach can handle qualitative as well as quantitative data, we consider both crisp and real-valued benchmark datasets from [15] and [24]. These datasets are medium-to-large in size, with between 32 and 699 objects per dataset and the number of conditional features ranging from 6 to 2556. All of them are consistent.

The quality of subsets found is evaluated using JRip [25] implemented in the WEKA toolkit [26]. JRip learns propositional rules by repeatedly growing rules and pruning them. During the growth phase, features are added greedily until a termination condition 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.

A. Crisp datasets

Table III contains general information about the used datasets, and shows for each approach the size of the resulting conditional feature set. Between brackets, we also list the γ value (i.e., the classical dependency degree) of this attribute subset. Due to Prop. 1, for $Q_{(0.2,1)}$, this value is always 1, but for the other approaches this does not hold in general. As can be seen, in general shorter or equal-length subsets are found with high dependency degrees. In some cases, like for $derm2$ with $\gamma^{Q_{(0,0.8)}}$ and $\gamma'^{Q_{(0.2,1)}}$, the results appear disappointing. Upon closer inspection, we noticed that in these cases QUICKREDUCT produced good-quality intermediary subsets (e.g. with $\gamma^{Q_{(0,0.8)}} \geq 0.95$) in early stages, but that afterwards the algorithm stalled by consecutively adding features with only a minimal increase of the dependency value. This might be mended by selecting a slightly lower α threshold to interrupt the process timely.

Table IV contains the accuracy results obtained when the dataset is reduced according to the found subset and classified with JRip using 10-fold cross validation. Again, comparable or better results are obtained in general with VQRS. It is also interesting to note that the "simplified" VQRS dependency degree γ'_{Q_u} does not produce worse results than the original version γ_{Q_u}, which is encouraging given the computational advantages of the former.

B. Real-valued datasets

Tables V and VI contain the experimental results for continuous decision systems. We used $\mathcal{T}_L$ and $\mathcal{I}_L$ as t-norm and implicator for defining fuzzy indiscernibility relations and for computing γ, respectively.

TABLE III
CRISP DATASET DETAILS

Dataset	Features	Objects	Subset size and γ value						
			γ	$\gamma^{Q_{\geq 0.9}}$	$\gamma^{Q_{(0,0.8)}}$	$\gamma^{Q_{(0.2,1)}}$	$\gamma'^{Q_{\geq 0.9}}$	$\gamma'^{Q_{(0,0.8)}}$	$\gamma'^{Q_{(0.2,1)}}$
breast	9	699	4	4(1.0)	5(0.98)	4(1.0)	4(1.0)	4(1.0)	4(1.0)
corral	6	64	5	5(1.0)	6(1.0)	6(1.0)	5(1.0)	5(1.0)	5(1.0)
derm	34	366	7	7(1.0)	7(1.0)	6(1.0)	7(1.0)	6(1.0)	6(1.0)
derm2	34	358	10	11(1.0)	26(1.0)	11(1.0)	11(1.0)	9(1.0)	25(1.0)
heart	13	294	7	7(1.0)	10(1.0)	8(1.0)	7(1.0)	7(1.0)	8(1.0)
ionos	34	230	8	7(0.93)	11(1.0)	9(1.0)	8(1.0)	8(1.0)	8(1.0)
lung	56	32	4	4(0.56)	3(0.19)	4(1.0)	5(1.0)	4(1.0)	4(1.0)
soybeanL	35	266	12	12(1.0)	13(1.0)	11(1.0)	12(1.0)	11(1.0)	11(1.0)
soybeanS	35	47	2	2(1.0)	2(1.0)	2(1.0)	2(1.0)	2(1.0)	2(1.0)
vote	16	300	9	16(1.0)	8(0.91)	16(1.0)	16(1.0)	9(1.0)	16(1.0)
water	38	521	15	13(1.0)	19(1.0)	13(1.0)	13(1.0)	14(1.0)	14(1.0)
zoo	16	101	5	6(1.0)	4(0.37)	5(1.0)	5(1.0)	5(1.0)	5(1.0)

TABLE IV
CRISP DATASET CLASSIFICATION ACCURACY

Dataset	JRip (%)						
	γ	$\gamma^{Q_{\geq 0.9}}$	$\gamma^{Q_{(0,0.8)}}$	$\gamma^{Q_{(0.2,1)}}$	$\gamma'^{Q_{\geq 0.9}}$	$\gamma'^{Q_{(0,0.8)}}$	$\gamma'^{Q_{(0.2,1)}}$
breast	95.6	95.6	94.3	94.7	95.6	95.6	95.6
corral	95.3	95.3	96.9	96.9	95.3	95.3	95.3
derm	76.2	86.1	80.9	80.9	86.1	80.9	70.5
derm2	89.7	90.8	93.9	86.6	90.8	91.1	93.0
heart	81.3	77.9	78.9	79.9	77.9	83	78.2
ionos	81.8	75.7	87.8	82.6	87.0	82.2	82.2
lung	84.4	84.4	71.9	84.4	84.4	87.5	84.4
soybeanL	84.2	84.2	82.3	83.9	83.1	83.1	83.1
soybeanS	100	100	100	100	100	100	100
vote	95.0	94.7	94.7	94.7	94.7	95.0	94.7
water	67.0	69.5	62.6	68.7	69.5	70.1	67.6
zoo	92.1	90.0	88.1	88.1	92.1	92.1	92.1

Similar observations can be made as for the crisp case. The benefit of VQRS is clearest for *web*, with considerably shorter subsets produced that achieve greater accuracy. Incidentally, note that for this dataset, especially remarkable because of its amount of features, VPRS performs worst. Also, it can be seen that there is no single best parameter combination for the fuzzy quantifier; apparently, this depends on the dataset, and on whether γ^{Q_u} or γ'^{Q_u} is used.

Finally, the hill-climbing style of the QUICKREDUCT heuristic may also limit the effectiveness of reduction. In future experiments, we plan to implement more advanced search algorithms, based e.g. on genetic algorithms or ant colony optimization, that can lead to more optimal subsets.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented a new fuzzy-rough approach to feature selection. Its key novelty is the incorporation of a more flexible lower approximation into the reduction process, which is motivated by the need to downsize the adverse effect of noise in datasets. As such, our framework bears much similarities to Ziarko's VPRS approach, which it generalizes by providing more, and smoother quantifiers. It also integrates noise-handling facilities with existing fuzzy-rough based approaches, and in particular with measures for approximate, rather than crisp, equality.

We conjecture that, with fuzzy quantifiers, the attribute selection procedure can be made more robust w.r.t. the quantifier parameters, but this point remains to be confirmed by further experimentation. From our preliminary observations, we noticed that the second parameter in $Q_{(\alpha,\beta)}$ is still quite determining for the quality of the resulting feature subsets, which also raises the possibility of learning optimal quantifiers from the data itself. This is an important avenue of further research.

We also noted that the VQRS and VPRS approaches satisfy less theoretical properties than their classical counterparts, but in practice this does not hamper the algorithms exceedingly. The non-monotonicity of γ^{Q_u} and γ'^{Q_u} remains a critical point, however, affecting e.g. the effectiveness of heuristic algorithms like QUICKREDUCT and REVERSEREDUCT. In [27], Ziarko recently proposed a

TABLE V

REAL-VALUED DATASET DETAILS

Dataset	Features	Objects	Subset size and γ value						
			γ	$\gamma^{Q_{\geq 0.9}}$	$\gamma^{Q_{(0,0.8)}}$	$\gamma^{Q_{(0.2,1)}}$	$\gamma'^{Q_{\geq 0.9}}$	$\gamma'^{Q_{(0,0.8)}}$	$\gamma'^{Q_{(0.2,1)}}$
cleveland	13	297	8	7(0.999)	8(1.0)	8(1.0)	8(1.0)	7(1.0)	9(1.0)
heart	13	270	7	7(0.999)	8(0.997)	8(1.0)	7(0.999)	7(1.0)	8(1.0)
ionosphere	34	230	8	8(1.0)	7(0.998)	7(1.0)	7(0.999)	6(0.998)	7(1.0)
olitos	25	120	5	5(1.0)	5(0.995)	5(1.0)	5(1.0)	5(0.993)	5(1.0)
water 2	38	390	6	6(0.999)	9(0.997)	6(1.0)	6(0.999)	6(0.996)	6(1.0)
water 3	38	390	6	6(1.0)	8(0.995)	7(1.0)	6(1.0)	5(0.996)	6(1.0)
web	2556	149	20	26(0.999)	12(0.749)	16(1.0)	27(1.0)	16(0.991)	16(1.0)
wine	13	178	5	5(1.0)	5(0.998)	5(1.0)	5(0.999)	5(1.0)	5(1.0)

TABLE VI

REAL-VALUED DATASET CLASSIFICATION ACCURACY

Dataset	JRip (%)						
	γ	$\gamma^{Q_{\geq 0.9}}$	$\gamma^{Q_{(0,0.8)}}$	$\gamma^{Q_{(0.2,1)}}$	$\gamma'^{Q_{\geq 0.9}}$	$\gamma'^{Q_{(0,0.8)}}$	$\gamma'^{Q_{(0.2,1)}}$
cleveland	54.5	55.2	54.2	54.5	52.9	55.6	53.5
heart	78.5	81.9	75.2	70	81.9	81.9	80.4
ionosphere	87.8	88.3	87.8	87.8	88.7	90.9	89.6
olitos	60.8	60.8	60.8	61.7	64.2	64.2	65.8
water 2	83.1	82.8	83.8	86.4	83.6	83.1	83.3
water 3	80.5	85.1	79.2	84.1	81.8	81.8	81.8
web	53.7	52.3	55.0	59.7	48.3	53.7	53
wine	95.5	92.7	95.5	90.4	95.5	93.3	95.5

monotonic dependency degree for VPRS-based attribute reduction in crisp datasets; an interesting challenge is therefore to extend this to the case of continuous data and fuzzy quantifiers.

REFERENCES

[1] L.A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, 338–353, 1965.

[2] Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, vol. 11(5), 341–356, 1982.

[3] D. Dubois, H. Prade, "Rough fuzzy sets and fuzzy rough sets," *International Journal of General Systems*, vol. 17, 91–209, 1990.

[4] C. Cornelis, M. De Cock, A. Radzikowska, "Fuzzy Rough Sets: from Theory into Practice," *Handbook of Granular Computing (W. Pedrycz, A. Skowron, V. Kreinovich, eds.)*, in press, 2008.

[5] A. Bargiela, W. Pedrycz, *Granular Computing. An introduction.* Kluwer Academic Publishers, 2002.

[6] L.A. Zadeh, "Soft Computing and Fuzzy Logic," *IEEE Software*, vol. 11(6), 48–56, 1994.

[7] A.M. Radzikowska, E.E. Kerre, E.E., "A comparative study of fuzzy rough sets," *Fuzzy Sets and Systems*, vol. 126, 137–156, 2002.

[8] C. Cornelis, M. De Cock and A. Radzikowska, "Vaguely Quantified Rough Sets,", *Proc. 11th Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC2007), Lecture Notes in Artificial Intelligence 4482*, 87–94, 2007.

[9] W. Ziarko, "Variable precision rough set model", *Journal of Computer and System Sciences*, vol. 46, 39-59, 1993.

[10] M. Dash, H. Liu, "Feature Selection for Classification", *Intelligent Data Analysis*, vol. 1(3), 131–156, 1997.

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

[12] Z. Pawlak, *Rough Sets — Theoretical aspects of reasoning about data.* Kluwer Academic Publishers, Dordrecht, Netherlands, 1991.

[13] H.S. Nguyen, "Discretization Problem for Rough Sets Methods", *1st Int. Conf. on Rough Sets and Current Trends in Computing (RSCTC'98)*, 545–552, 1198.

[14] J.W. Grzymala-Busse, J. Stefanowski, "Three discretization methods for rule induction", *International Journal of Intelligent Systems*, vol. 16(1), 29-38 (2001).

[15] R. Jensen, Q. Shen, "Fuzzy-rough sets assisted attribute selection," *IEEE Transactions on Fuzzy Systems*, vol. 15(1), 73–89, 2007.

[16] R. Jensen, Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Transactions on Fuzzy Systems*, to appear, 2008.

[17] W. Ziarko, "Decision Making with Probabilistic Decision Tables", *Proc. 7th Int. Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing (RSFDGrC'99)* , 463-471, 1999.

[18] C. Cornelis, G. Hurtado Martín, R. Jensen, D. Ślęzak, "Feature Selection with Fuzzy Decision Reducts," *3rd Int. Conf. on Rough Sets and Knowledge Technology (RSKT'08)*, submitted, 2008.

[19] P. Smets, P. Magrez, "Implication in fuzzy logic," *International Journal of Approximate Reasoning*, vol. 1(4) 327–347, 1987.

[20] L.A. Zadeh, "A Computational Approach to Fuzzy Quantifiers in Natural Languages," *Computers and Mathematics with Applications*, Vol. 9, 149–184, 1983.

[21] W. Ziarko, "Set approximation quality measures in the variable precision rough set model," *Soft Computing Systems: Design, Management and Applications (A. Abraham, J. Ruiz-del-Solar, M. Koppen, eds.)*, IOS Press, 442–452, 2002.

[22] A. Chouchoulas, Q. Shen, "Rough Set-Aided Keyword Reduction for Text Categorisation", *Applied Artificial Intelligence*, vol. 15(9), 843–873, 2001.

[23] M. De Cock, E.E. Kerre, "On (Un)suitable Fuzzy Relations to Model Approximate Equality", *Fuzzy Sets and Systems*, vol. 133(2), 137–153, 2003.

[24] C.L. Blake, C.J. Merz, *UCI Repository of Machine Learning Databases.* Irvine, University of California, 1998. http://www.ics.uci.edu/˜mlearn/

[25] W.W. Cohen, "Fast Effective Rule Induction," *Proc. 12th Int. Conf. on Machine Learning*, 115–123, 1995.

[26] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.

[27] W. Ziarko, "Probabilistic Approach to Rough Sets", *International Journal of Approximate Reasoning*, in press.