Fuzzy-Rough Sets Assisted Attribute Selection

Shen, Qiang; Jensen, Richard

Published in:
IEEE Transactions on Fuzzy Systems

DOI:
10.1109/TFUZZ.2006.889761

Publication date:
2007

Citation for published version (APA):
Shen, Q., & Jensen, R. (2007). Fuzzy-Rough Sets Assisted Attribute Selection. *IEEE Transactions on Fuzzy Systems*, *15*(1), 73-89. https://doi.org/10.1109/TFUZZ.2006.889761

Fuzzy-Rough Sets Assisted Attribute Selection

Richard Jensen and Qiang Shen

***Abstract*—Attribute selection (AS) refers to the problem of selecting those input attributes or features that are most predictive of a given outcome; a problem encountered in many areas such as machine learning, pattern recognition and signal processing. Unlike other dimensionality reduction methods, attribute selectors preserve the original meaning of the attributes after reduction. This has found application in tasks that involve datasets containing huge numbers of attributes (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. AS techniques have also been applied to small and medium-sized datasets in order to locate the most informative attributes for later use. One of the many successful applications of rough set theory has been to this area. The rough set ideology of using only the supplied data and no other information has many benefits in AS, where most other methods require supplementary knowledge. However, the main limitation of rough set-based attribute selection in the literature is the restrictive requirement that all data is discrete. In classical rough set theory, it is not possible to consider real-valued or noisy data. This paper investigates a novel approach based on fuzzy-rough sets, fuzzy rough feature selection (FRFS), that addresses these problems and retains dataset semantics. FRFS is applied to two challenging domains where a feature reducing step is important; namely, web content classification and complex systems monitoring. The utility of this approach is demonstrated and is compared empirically with several dimensionality reducers. In the experimental studies, FRFS is shown to equal or improve classification accuracy when compared to the results from unreduced data. Classifiers that use a lower dimensional set of attributes which are retained by fuzzy-rough reduction outperform those that employ more attributes returned by the existing crisp rough reduction method. In addition, it is shown that FRFS is more powerful than the other AS techniques in the comparative study.**

***Index Terms*—Attribute selection, dimensionality reduction, fuzzy-rough sets, rough selection.**

I. Introduction

THERE are many factors that motivate the inclusion of a dimensionality reduction (DR) step in a variety of problem-solving systems [5]. Many application problems process data in the form of a collection of real-valued vectors (for example, text classification [45], bookmark categorization [15]). If these vectors exhibit a high dimensionality, then processing becomes infeasible. Therefore, it is often useful, and sometimes necessary, to reduce the data dimensionality to a more manageable size with as little information loss as possible.

Sometimes, high-dimensional complex phenomena can be governed by significantly fewer, simple variables [11]. The process of dimensionality reduction here will act as a tool for modelling these phenomena, improving their clarity. There is often a significant amount of redundant or misleading information present; this will need to be removed before any further processing can be carried out. For example, the problem of deriving classification rules from large datasets often benefits from a data reduction preprocessing step [33]. Not only does this reduce the time required to perform induction, but it makes the resulting rules more comprehensible and can increase the resulting classification accuracy.

Whereas semantics-destroying dimensionality reduction techniques irreversibly transform data, semantics-preserving DR techniques (referred to as attribute selection) attempt to retain the meaning of the original attribute set. The main aim of attribute selection is to determine a minimal attribute subset from a problem domain while retaining a suitably high accuracy in representing the original attributes.

There are often many attributes involved, and combinatorially large numbers of attribute combinations, to select from. Note that the number of attribute subset combinations with m attributes from a collection of N total attributes is $N!/[m!(N-m)!]$. It might be expected that the inclusion of an increasing number of attributes would increase the likelihood of including enough information to distinguish between classes. Unfortunately, this is not true if the size of the training dataset does not also increase rapidly with each additional attribute included. This is the so-called curse of dimensionality [3]. A high-dimensional dataset increases the chances that a data-mining algorithm will find spurious patterns that are not valid in general. 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.

A technique that can reduce dimensionality using information contained within the data set and that preserves the meaning of the attributes (i.e., semantics-preserving) is clearly desirable. Rough set theory (RST) can be used as such a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information [23], [26].

Over the past ten years, RST has indeed become a topic of great interest to researchers and has been applied to many domains [16]. Given a dataset with discretized attribute values, it is possible to find a subset (termed a *reduct*) of the original attributes using RST that are the most informative; all other attributes can be removed from the dataset with minimal information loss. From the dimensionality reduction perspective, informative attributes are those that are most predictive of the class attribute.

However, it is most often the case that the values of attributes may be both crisp and *real-valued*, and this is where traditional rough set theory encounters a problem. It is not possible in the original theory to say whether two attribute values are similar

Manuscript received February 1, 2005; revised December 18, 2005 and June 16, 2006.

The authors are with the Department of Computer Science, The University of Wales, Aberystwyth, Ceredigion SY23 3DB, Wales, U.K. (e-mail: rkj@aber.ac.uk; qqs@aber.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2006.889761

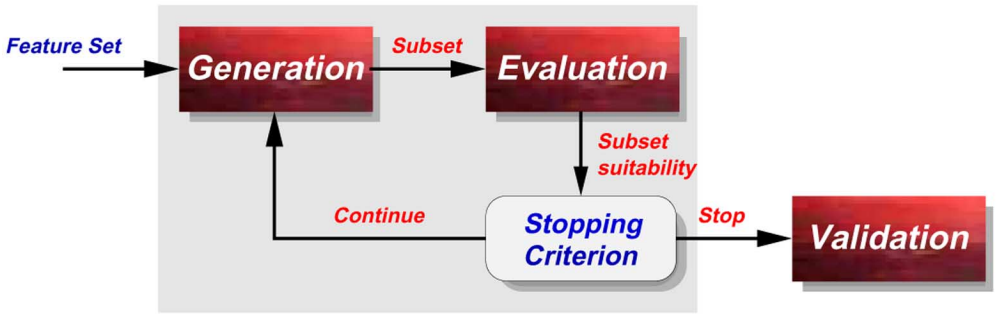

Fig. 1. Attribute selection process.

and to what extent they are the same; for example, two close values may only differ as a result of noise, but in RST they are considered to be as different as two values of a different order of magnitude. As a result of this, extensions to the original theory have been proposed, for example those based on similarity or tolerance relations [36], [38], [39].

It is, therefore, desirable to develop techniques to provide the means of data reduction for crisp and real-value attributed datasets which utilizes the extent to which values are similar. This can be achieved through the use of *fuzzy-rough* sets. Fuzzy-rough sets encapsulate the related but distinct concepts of vagueness (for fuzzy sets [46]) and indiscernibility (for rough sets), both of which occur as a result of uncertainty in knowledge [9]. Vagueness arises due to a lack of sharp distinctions or boundaries in the data itself. This is typical of human communication and reasoning.

This paper presents a method, fuzzy-rough feature selection (FRFS), that employs fuzzy-rough sets to provide 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 attribute which can be automatically derived from the data. In the work presented here, *all* fuzzy sets are derived solely from the data. This avoids the need for domain experts to provide information on the data involved and ties in with the advantage of rough sets in that it requires no information other than the data itself. However, if such experts are readily available, it is beneficial to capture their knowledge in the form of fuzzy data partitions to improve the transparency of the selection process and any other future processes (e.g., rule induction).

The rest of this paper is structured as follows. An introduction to the attribute selection problem is presented in Section II, describing the main components of an attribute selector. Section III introduces the theory main concepts behind crisp rough set=based attribute reduction. Next, the fuzzy-rough set-based attribute selection method is described in detail. The new fuzzy-rough attribute evaluation metric is compared with several of the leading metrics using artificial data. FRFS is then applied to two challenging areas: website categorization and complex systems monitoring. The paper is concluded in Section VIII.

II. Attribute Selection

Semantics-preserving dimensionality reduction techniques attempt to retain the meaning of the original attribute set. The main aim of attribute selection is to determine a minimal attribute subset from a problem domain while retaining a suitably high accuracy in representing the original attributes. In many real world problems, AS is a must due to the abundance of noisy, irrelevant or misleading attributes. A detailed review of attribute selection techniques devised for classification tasks can be found in [8].

The usefulness of an attribute or attribute subset is determined by both its *relevancy* and *redundancy*. An attribute is said to be relevant if it is predictive of the decision attribute(s), otherwise it is irrelevant. An attribute is considered to be redundant if it is highly correlated with other attributes. Hence, the search for a good attribute subset involves finding those attributes that are highly correlated with the decision attribute(s), but are uncorrelated with each other.

Given an attribute set size n, the task of AS can be seen as a search for an "optimal" attribute subset through the competing 2^n candidate subsets. The definition of what an optimal subset is may vary depending on the problem to be solved. Although an exhaustive method may be used for this purpose in theory, this is quite impractical for most datasets. Usually AS algorithms involve heuristic or random search strategies in an attempt to avoid this prohibitive complexity. However, the degree of optimality of the final attribute subset is often reduced. The overall procedure for any attribute selection method is given in Fig. 1 [8].

The generation procedure implements a search method [19], [35] that generates subsets of attributes for evaluation. It may start with no attributes, all attributes, a selected attribute set or some random attribute subset. Those methods that start with an initial subset usually select these attributes heuristically beforehand. Attributes are added (*forward selection*) or removed (*backward elimination*) iteratively in the first two cases [8]. In the last case, attributes are either iteratively added or removed or produced randomly thereafter. An alternative selection strategy is to select instances and examine differences in their attributes. The evaluation function calculates the suitability of an attribute subset produced by the generation procedure and compares this with the previous best candidate, replacing it if found to be better.

A stopping criterion is tested every iteration to determine whether the AS process should continue or not. For example, such a criterion may be to halt the AS process when a certain number of attributes have been selected if based on the generation process. A typical stopping criterion centred on the evaluation procedure is to halt the process when an optimal subset is reached. Once the stopping criterion has been satisfied, the loop terminates. For use, the resulting subset of attributes may be validated.

Determining subset optimality is a challenging problem. There is always a trade-off in non-exhaustive techniques between subset minimality and subset suitability—the task is to decide which of these must suffer in order to benefit the other. For some domains (particularly where it is costly or impractical to monitor many attributes), it is much more desirable to have a smaller, less accurate attribute subset. In other areas it may be the case that the modelling accuracy (e.g., the classification rate) using the selected attributes must be extremely high, at the expense of a non-minimal set of attributes.

Attribute selection algorithms may be classified into two categories based on their evaluation procedure. If an algorithm performs AS independently of any learning algorithm (i.e., it is a completely separate preprocessor), then it is a *filter* approach. In effect, irrelevant attributes are filtered out before induction. Filters tend to be applicable to most domains as they are not tied to any particular induction algorithm.

If the evaluation procedure is tied to the task (e.g., classification) of the learning algorithm, the AS algorithm employs the *wrapper* approach. This method searches through the attribute subset space using the estimated accuracy from an induction algorithm as a measure of subset suitability. Although wrappers may produce better results, they are expensive to run and can break down with very large numbers of attributes. This is due to the use of learning algorithms in the evaluation of subsets, some of which can encounter problems when dealing with large datasets.

III. Rough Set Attribute Reduction

RSAR [7] 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 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.

To illustrate the operation of these, an example dataset (Table I) will be used. Here, the table consists of four conditional attributes (a, b, c, d), one decision attribute (e) and eight objects. The task of attribute selection here is to choose the smallest subset of these conditional attributes so that the resulting reduced dataset remains consistent with respect to the decision attribute.

TABLE I
Example Dataset

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

A. Theoretical Background

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) and $\mathbb{A}$ is a nonempty 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 $\mathrm{IND}(P)$

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

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

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

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

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \oslash\}. \quad (3)$$

If $(x, y) \in \mathrm{IND}(P)$, 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. $\mathrm{IND}(P)$ creates the following partition of $\mathbb{U}$:

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

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

$$\underline{P}X = \{x \,|\, [x]_P \subseteq X\} \quad (4)$$

$$\overline{P}X = \{x \,|\, [x]_P \cap X \neq \oslash\}. \quad (5)$$

Let P and Q be equivalence relations over $\mathbb{U}$, then the positive, negative, and boundary regions can be defined as

$$\begin{aligned}\mathrm{POS}_P(Q) &= \bigcup_{X\in\mathbb{U}/Q} \underline{P}X \\ \mathrm{NEG}_P(Q) &= \mathbb{U} - \bigcup_{X\in\mathbb{U}/Q} \overline{P}X \\ \mathrm{BND}_P(Q) &= \bigcup_{X\in\mathbb{U}/Q} \overline{P}X - \bigcup_{X\in\mathbb{U}/Q} \underline{P}X.\end{aligned}$$

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, $\mathrm{BND}_P(Q)$, is the set of objects that can possibly, but not certainly, be classified in this way. The negative region, $\mathrm{NEG}_P(Q)$, 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

$$\begin{aligned}\mathrm{POS}_P(Q) &= \bigcup\{\oslash, \{2,5\}, \{3\}\} = \{2,3,5\} \\ \mathrm{NEG}_P(Q) &= \mathbb{U} - \bigcup\{\{0,4\}, \{2,0,4,1,6,7,5\}, \{3,1,6,7\}\} \\ &= \oslash \\ \mathrm{BND}_P(Q) &= \mathbb{U} - \{2,3,5\} = \{0,1,4,6,7\}.\end{aligned}$$

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

$$k = \gamma_P(Q) = \frac{|\mathrm{POS}_P(Q)|}{|\mathbb{U}|} \tag{6}$$

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 $\{e\}$ from the attributes $\{b, c\}$ is:

$$\begin{aligned}\gamma_{\{b,c\}}(\{e\}) &= \frac{|\mathrm{POS}_{\{b,c\}}(\{e\})|}{|\mathbb{U}|} \\ &= \frac{|\{2,3,5\}|}{|\{0,1,2,3,4,5,6,7\}|} = \frac{3}{8}.\end{aligned}$$

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

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

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

$$\begin{aligned}\gamma_{\{a,b,c\}}(\{e\}) &= |\{2,3,5,6\}|/8 = 4/8 \\ \gamma_{\{a,b\}}(\{e\}) &= |\{2,3,5,6\}|/8 = 4/8 \\ \gamma_{\{b,c\}}(\{e\}) &= |\{2,3,5\}|/8 = 3/8 \\ \gamma_{\{a,c\}}(\{e\}) &= |\{2,3,5,6\}|/8 = 4/8\end{aligned}$$

and calculating the significance of the three attributes gives

$$\begin{aligned}\sigma_P(Q, a) &= \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{b,c\}}(\{e\}) = 1/8 \\ \sigma_P(Q, b) &= \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,c\}}(\{e\}) = 0 \\ \sigma_P(Q, c) &= \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,b\}}(\{e\}) = 0.\end{aligned}$$

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

B. 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* is defined as a subset of minimal cardinality $R_{\min}$ of the conditional attribute set $\mathbb{C}$ such that $\gamma_{R_{\min}}(\mathbb{D}) = \gamma_{\mathbb{C}}(\mathbb{D})$

$$R = \{X : X \subseteq \mathbb{C}, \gamma_X(\mathbb{D}) = \gamma_{\mathbb{C}}(\mathbb{D})\} \tag{8}$$

$$R_{\min} = \{X : X \in R, \forall Y \in R, |X| \leq |Y|\}. \tag{9}$$

The intersection of all the sets in $R_{\min}$ 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

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

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

$$R_{\min} = \{\{b, d\}, \{c, d\}\}$$

If $\{b, d\}$ is chosen, then the dataset can be reduced as in Table II. Clearly, each object can be uniquely classified according to the attribute values remaining.

TABLE II
REDUCED DATASET

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

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

Fig. 2. QUICKREDUCT algorithm.

The problem of finding a reduct of an information system has been the subject of much research [2], [40]. The most basic solution to locating such a subset is to simply generate *all* possible subsets and retrieve those with a maximum rough set dependency degree. Obviously, this is an expensive solution to the problem and is only practical for very simple datasets. Most of the time only one reduct is required as, typically, only one subset of attributes is used to reduce a dataset, so all the calculations involved in discovering the rest are pointless.

To improve the performance of the above method, an element of pruning can be introduced. By noting the cardinality of any prediscovered reducts, the current possible subset can be ignored if it contains more elements. However, a better approach is needed—one that will avoid wasted computational effort.

The QUICKREDUCT algorithm given in Fig. 2 (adapted from [7]), calculates 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. Other such techniques may be found in [25].

According to the QUICKREDUCT algorithm, the dependency of each attribute is calculated, and the best candidate chosen. In Fig. 3, this stage is illustrated using the example dataset. As attribute d generates the highest dependency degree, then that attribute is chosen and the sets $\{a, d\}$, $\{b, d\}$ and $\{c, d\}$ are evaluated. This process continues until the dependency of the reduct 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 most datasets. However, it may be infeasible for very large data, so alternative stopping criteria may have to be used.

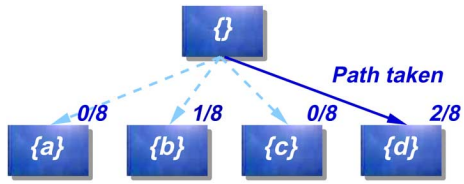

Fig. 3. Branches of the search space.

One such criterion could be to terminate the search when there is no further increase in the dependency measure. This will produce exactly the same path to a reduct due to the monotonicity of the measure [7], without the computational overhead of calculating the dataset consistency.

Other developments include REVERSEREDUCT where the strategy is backward elimination of attributes as opposed to the current forward selection process. Initially, all attributes appear in the reduct candidate; the least informative ones are incrementally removed until no further attribute can be eliminated without introducing inconsistencies. This is not often used for large datasets, as the algorithm must evaluate large attribute subsets (starting with the set containing *all* attributes) which is too costly, although the computational complexity is, in theory, the same as that of forward-looking QUICKREDUCT. As both forward and backward methods perform well, it is thought that a combination of these within one algorithm would be effective.

This, however, is not guaranteed to find a *minimal* reduct. Using the dependency function to discriminate between candidates may lead the search down a nonminimal 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.

It is interesting to note that the rough set degree of dependency measure is very similar to the consistency criterion used by the FOCUS algorithm and others [1], [31]. In FOCUS, a breadth-first search is employed such that any subset is rejected if this produces at least one inconsistency. If this is converted into a guided search using the consistency measure as a heuristic, it should behave exactly as QUICKREDUCT. Consistency is defined as the number of discernible objects out of the entire object set—exactly that of the dependency measure.

IV. 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 [33], 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 fuzzy 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], [22], it is possible to use this information to better guide attribute selection.

A. 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], [41], [44]. 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 transitivity ($\mu_S(x,z)\geq \mu_S(x,y)\wedge\mu_S(y,z)$) 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.

B. Fuzzy-Rough Sets

From the literature, the fuzzy P-lower and P-upper approximations are defined as [9]

$$\mu_{\underline{P}X}(F_i)=\inf_x \max\{1-\mu_{F_i}(x),\mu_X(x)\} \quad \forall i \tag{10}$$

$$\mu_{\overline{P}X}(F_i)=\sup_x \min\{\mu_{F_i}(x),\mu_X(x)\} \quad \forall i \tag{11}$$

where P is an attribute subset, X is the concept to be approximated, and F_i denotes a fuzzy equivalence class belonging to $\mathbb{U}/P$. Note that although the universe of discourse in attribute selection is finite, this is not the case in general, hence the use of $\sup$ and $\inf$. These definitions diverge a little from the crisp upper and lower approximations, as the memberships of individual objects to the approximations are not explicitly available. As a result of this, the fuzzy lower and upper approximations are herein redefined as

$$\mu_{\underline{P}X}(x) = \sup_{F\in\mathbb{U}/P} \min(\mu_F(x), \inf_{y\in\mathbb{U}} \max\{1-\mu_F(y),\mu_X(y)\}) \tag{12}$$

$$\mu_{\overline{P}X}(x) = \sup_{F\in\mathbb{U}/P} \min(\mu_F(x), \sup_{y\in\mathbb{U}} \min\{\mu_F(y),\mu_X(y)\}). \tag{13}$$

In the implementation of the fuzzy-rough reduction method, not all $y\in\mathbb{U}$ need to be considered—only those where $\mu_F(y)$ is nonzero, i.e., where object y is a fuzzy member of (fuzzy) equivalence class F. The tuple $\langle \underline{P}X, \overline{P}X\rangle$ is called a *fuzzy-rough set*. It can be seen that these definitions degenerate to traditional rough sets when all equivalence classes are crisp [15].

C. Fuzzy-Rough Reduction Process

Fuzzy-rough set-based attribute 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 [47], the membership of an object $x\in\mathbb{U}$, belonging to the fuzzy positive region can be defined by

$$\mu_{\mathrm{POS}_P(Q)}(x)=\sup_{X\in\mathbb{U}/Q}\mu_{\underline{P}X}(x). \tag{14}$$

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:

$$\gamma'_P(Q)=\frac{\left|\mu_{\mathrm{POS}_P(Q)}(x)\right|}{|\mathbb{U}|}=\frac{\sum_{x\in\mathbb{U}}\mu_{\mathrm{POS}_P(Q)}(x)}{|\mathbb{U}|}. \tag{15}$$

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 $\mu_{\mathrm{POS}_P(Q)}(x)$ 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 $\mathbb{U}/\mathrm{IND}(\{a\})$ and $\mathbb{U}/\mathrm{IND}(\{b\})$ must be considered in determining $\mathbb{U}/P$. In general

$$\mathbb{U}/P=\otimes\{a\in P:\mathbb{U}/\mathrm{IND}(\{a\})\} \tag{16}$$

where

$$A\otimes B=\{X\cap Y:\forall X\in A,\forall Y\in B, X\cap Y\neq\oslash\}. \tag{17}$$

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 $F_i, i=1,2,\ldots,n$

$$\mu_{F_1\cap\ldots\cap F_n}(x)=\min(\mu_{F_1}(x),\mu_{F_2}(x),\ldots,\mu_{F_n}(x)). \tag{18}$$

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) $\forall x \in (\mathbb{C} - R)$
(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. 4. Fuzzy-rough QUICKREDUCT algorithm.

D. 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.

A possible way of combatting this would be to determine the degree of dependency of a set of decision attributes D upon the full attribute set and use this as the denominator rather than $|\mathbb{U}|$ (for normalization), allowing γ' to reach 1. With these issues in mind, a fuzzy-rough hill-climbing search algorithm has been developed as given in Fig. 4. 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 the fuzzy-rough degree of dependency measure is nonmonotonic, 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. This provided the motivation for the use of an alternative search mechanism based on ant colony optimization [17]. However, the algorithm as presented in Fig. 4 is still highly useful in locating good subsets quickly.

It is also possible to reverse the search process in a manner identical to that of REVERSEREDUCT; that is, start with the full set of attributes and incrementally remove the least informative attributes. This process continues until no more attributes can be removed without reducing the total number of discernible objects in the dataset. Again, this tends not to be applied to larger datasets as the cost of evaluating these larger attribute subsets is too great.

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.

TABLE III
EXPERIMENTAL COMPARISON OF THE TWO FORMULATIONS FOR THE CALCULATION OF THE POSITIVE REGION

Dataset	No. of Attributes	Eq. (19) (s)	Eq. (20) (s)	Opt. (s)
Glass	10	29.5	26.7	7.18
Wine	14	5.41	3.05	2.20
Olitos	26	47.6	21.9	13.0
JobSat	27	19.2	5.75	2.72
Ionosphere	35	204.5	107.8	76.9
Selwood	54	57.5	15.9	5.64
Isolet	618	368.4	131.9	47.2
Phenetyl	629	740.7	145.0	70.3
Caco	714	2709.3	213.5	114.1

E. Optimizing FRFS

There are several optimizations that can be implemented to speed up the FRFS process. The original definition of the fuzzy positive region, given in (14), can be more explicitly defined as

$$\mu_{\mathrm{POS}_P(Q)}(x) = \sup_{X\in\mathbb{U}/Q} \{ \sup_{F\in\mathbb{U}/P} \min(\mu_F(x) \inf_{y\in\mathbb{U}} \max\{1-\mu_F(y), \mu_X(y)\})\} \quad (19)$$

where P is a subset of the conditional attributes, Q the decision attribute(s). In order to speed up computation time, (19) can be rewritten as

$$\mu_{\mathrm{POS}_P(Q)}(x) = \sup_{F\in\mathbb{U}/P} \{\min(\mu_F(x) \sup_{X\in\mathbb{U}/Q} \{\inf_{y\in\mathbb{U}} \max(1-\mu_F(y), \mu_X(y))\})\}. \quad (20)$$

This reformulation helps to speed up the calculation of the fuzzy positive region by considering each fuzzy equivalence class F in U/P first. If the object x is found not to belong to F, the remainder of the calculations for this class need not be evaluated, due to the use of the min operator. This can save substantial time, as demonstrated in Table III, where the two definitions of the positive region are used to determine reducts from several small to large datasets. The times here are the times taken for each version of FRFS to find a reduct. Each version of FRFS will follow exactly the same route and will locate identical reducts. All the datasets are from the machine learning repository [4] and contain real-valued conditional attributes with nominal classifications.

Additionally in Table III, average runtimes are given for the optimized implementation of the fuzzy-rough attribute selector. This includes the use of the algorithm presented in Fig. 5, which is designed to result in faster computation of the fuzzy-rough metric for small attribute subsets. Excess computation is avoided at lines (4) and (6) which exploit the nature of t-norms and s-norms in the definitions of the lower approximation and positive region.

V. EVALUATING THE FUZZY-ROUGH METRIC

In order to evaluate the utility of the new fuzzy-rough measure of attribute significance, a series of artificial datasets were generated and used for comparison with five other leading attribute ranking measures. The datasets were created by generating around 30 random attribute values for 400 objects. Two or three attributes (referred to as x, y, or z) are chosen to contribute to the final boolean classification by means of an inequality. The

CALCULATEGAMMA'($\mathbb{P}$,$\mathbb{Q}$).
$\mathbb{P}$, the current attribute subset;
$\mathbb{Q}$, the set of decision attributes.

(1) $mems[], \gamma' \leftarrow 0$
(2) $\forall\, F \in \mathbb{U}/P$
(3) $deg = \sup_{X \in \mathbb{U}/Q} (\{\inf_{y \in \mathbb{U}} max\{1 - \mu_F(y), \mu_X(y)\}\}$
(4) **if** $deg \neq 0$
(5) $\forall o \in \mathbb{U}$
(6) **if** $mems[o] \neq 1$ && $deg > mems[o]$
(7) $mems[o] = max(min(\mu_F(o), deg), mems[o])$
(8) $\forall o \in \mathbb{U}$, γ' += $mems[o]$
(9) **return** γ'

Fig. 5. Optimized γ' calculation for small subsets.

task for the attribute rankers was to discover those attributes that are involved in the inequalities, ideally rating the other irrelevant attributes poorly in contrast.

A. Compared Metrics

The metrics compared are: the fuzzy-rough measure (FR), relief-F (Re), information gain (IG), gain ratio (GR), OneR (1R) and the statistical measure χ^2. The implementation of these metrics, apart from the fuzzy-rough measure, was obtained from [42]. A brief description of each is presented next.

1) Information Gain: The IG [13] is the expected reduction in entropy resulting from partitioning the dataset objects according to a particular attribute. The entropy of a labelled collection of objects S is defined as

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{21}$$

where p_i is the probability, typically approximated by the proportion of s belonging to class i. Based on this, the IG metric is

$$\text{IG}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \tag{22}$$

where $\text{values}(A)$ is the set of values for attribute A, S the set of training examples, S_v the set of training objects where A has the value v.

2) Gain Ratio: One limitation of the IG measure is that it favours attributes with many values. The GR seeks to avoid this bias by incorporating another term, split information, that is sensitive to how broadly and uniformly the attribute splits the considered data

$$\text{Split}(S, A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \tag{23}$$

where each S_i is a subset of objects generated by partitioning S with the c-valued attribute A. The GR is then defined as follows:

$$\text{GR}(S, A) = \frac{IG(S, A)}{\text{Split}(S, A)}. \tag{24}$$

3) χ^2 Measure: In the χ^2 method [14], attributes are individually evaluated according to their χ^2 statistic with respect to the classes. For a numeric attribute, the method first requires its range to be discretized into several intervals. The χ^2 value of an attribute is defined as

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{25}$$

where m is the number of intervals; k the number of classes, A_{ij} the number of samples in the ith interval, jth class; R_i the number of objects in the ith interval; C_j the number of objects in the jth class; N the total number of objects; and E_{ij} the expected frequency of $A_{ij} (E_{ij} = R_i * C_j / N)$. The larger the χ^2 value, the more important the attribute.

4) Relief-F: Relief [18] evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. The distance between two objects is the sum of the number of attributes that differ in value between them, for nominal values. When dealing with continuous attributes, the distance is the normalised sum of the difference in attribute values. Relief-F extends this idea to dealing with multiclass problems as well as handling noisy and incomplete data. When used for attribute selection, the user must supply a threshold which determines the level of relevance that attributes must surpass in order to be finally chosen.

5) OneR: The OneR classifier [12] learns a one-level decision tree, i.e., it generates a set of rules that test one particular attribute. One branch is assigned for every value of an attribute; each branch is assigned the most frequent class. The error rate is then defined as the proportion of instances that do not belong to the majority class of their corresponding branch. Attributes with the higher classification rates are considered to be more significant than those resulting in lower accuracies.

B. Metric Comparison

The tables presented in this section show the results for the application of the metrics (outlined in Section V-A above) to the artificial data. The task for these metrics is to detect those attributes appearing in the datasets that affect the classifications. A good metric must also ignore attributes that are irrelevant, i.e., have no bearing upon the classification. The final row in each table indicates whether all irrelevant attributes are given a ranking of zero. The full results can be seen in Tables XVI–XXI.

For the data presented in Table IV, the first attribute, x, is used to determine the classification. The values of attributes y and z are derived from x: $y = \sqrt{x}, z = x^2$. Hence, a good feature ranker should detect the importance of these attributes, and consider all remaining attributes as irrelevant. It can be observed from the table that all metrics successfully rank the influential attributes highest. IG, GR, 1R, and χ^2 rank these attributes equally, whereas Re and FR rank attribute z higher. Only FR, IG, GR, and χ^2 rate all the other attributes as zero.

Thus, attribute rankers can discover the influential attributes but on their own are incapable of determining multiple attribute

TABLE IV
ATTRIBUTE EVALUATION FOR $x > 0.5, y = \sqrt{x}, z = x^2$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.5257	0.31758	0.997	1.0	99.5	200
y	0.5296	0.24586	0.997	1.0	99.5	200
z	0.5809	0.32121	0.997	1.0	99.5	200
others	= 0	$\neq 0$	= 0	= 0	$\neq 0$	= 0

TABLE V
ATTRIBUTE EVALUATION FOR $(x + y)^2 > 0.25$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.2330	0.1862	0.2328	0.1579	86.75	128.47
y	0.2597	0.1537	0.1687	0.169	87.75	71.97
others	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$

TABLE VI
ATTRIBUTE EVALUATION FOR $(x + y)^2 > 0.5$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.209	0.140067	0.241	0.156	79	119.56
y	0.2456	0.151114	0.248	0.165	78.25	122.34
others	= 0	$\neq 0$	= 0	= 0	$\neq 0$	= 0

TABLE VII
ATTRIBUTE EVALUATION FOR $(x + y)^3 < 0.125$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.2445	0.1486	0.134	0.134	87.75	57.46
y	0.2441	0.1659	0.159	0.164	87.25	73.39
others	= 0	$\neq 0$	= 0	= 0	$\neq 0$	= 0

interactions. Table IV could be reduced to one attribute only (either x, y, or z) without any loss of information as only these contribute to the classification. However, the rankers all rate these attributes highly and would only provide enough information to reduce the data to at least these three attributes. Here, the rankers have found the predictive (or relevant) attributes but have been unable to determine which of these are redundant.

Table V shows the results for the inequality $(x+y)^2 > 0.25$. If this inequality holds for an object then it is classified as 1, with a classification of 0 otherwise. Hence, both attributes x and y are required for deciding the classification. All attribute rankers evaluated detect this. FR, IG, GR, 1R and χ^2 also rank the tenth attribute highly—probably due to a chance correlation with the decision. The results in Table VI are for a similar inequality, with all the attribute rankers correctly rating the important attributes. FR, IG, GR, and χ^2 evaluate the remaining attributes as having zero significance.

In Table VII, all metrics apart from 1R locate the relevant attributes. For this dataset, 1R chooses 22 attributes as being the most significant, whilst ranking attributes x and y last. This may be due to the discretization process that must precede the application of 1R. If the discretization is poor, then the resulting attribute evaluations will be affected.

Table VIII shows the results for data classified by $x * y * z > 0.125$. All attribute rankers correctly detect these variables. However, in Table IX the results can be seen for the same inequality but with the impact of variable z increased. All metrics determine that z has the most influence on the decision, and almost all choose x and y next. Again, the 1R measure fails and chooses attributes 15, 19 and 24 instead.

TABLE VIII
ATTRIBUTE EVALUATION FOR $x * y * z > 0.125$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.1057	0.0750547	0.169	0.123	64.25	73.65
y	0.0591	0.1079423	0.202	0.226	66.75	88.04
z	0.1062	0.0955878	0.202	0.16	67.5	84.28
others	= 0	$\neq 0$	= 0	= 0	$\neq 0$	= 0

TABLE IX
ATTRIBUTE EVALUATION FOR $x * y * z^2 > 0.125$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.1511	0.098	0.1451	0.0947	76.5	65.43
y	0.1101	0.05571	0.0909	0.108	78.0	35.36
z	0.2445	0.14736	0.2266	0.2271	79.75	93.81
others	= 0	$\neq 0$	= 0	= 0	$\neq 0$	= 0

In summary, only the FR and Re metrics are applicable to datasets where the decision attribute is continuous. Both methods find the attributes that are involved in generating the decision values. This short investigation into the utility of the new fuzzy-rough measure has shown that it is comparable with the leading measures of attribute importance. Indeed, its behaviour is quite similar to the information gain and gain ratio metrics. This is interesting as both of these measures are entropy-based: an attribute subset with a maximum (crisp) rough set dependency has a corresponding entropy of 0. Unlike these metrics, the fuzzy-rough measure may also be applied to datasets containing real-valued decision attributes.

VI. APPLICATION TO WEBSITE CATEGORIZATION

There are an estimated 1 billion web pages available on the Internet with around 1.5 million web pages being added every day. The task to find a particular web page, which satisfies a user's requirements by traversing hyper-links, is very difficult. To aid this process, many web directories have been developed—some rely on manual categorization whilst others make decisions automatically. However, as web page content is vast and dynamic, manual categorization is becoming increasingly impractical. Automatic web site categorization is therefore required to deal with these problems.

The keywords extracted from web pages are weighted not only according to their statistical occurrence but also to their location within the document itself. These weights are almost always real-valued, which can be a problem for most attribute selectors unless data discretisation takes place (a source of information loss). This motivates the application of FRFS to this domain.

A key issue in the design of the system was that of modularity; it should be modelled in such a way as to enable the straightforward replacement of existing techniques with new methods. The current implementation allows this flexibility by dividing the overall process into several independent sub-modules (see Fig. 6).

The training and testing datasets were generated using Yahoo [43]. Five classification categories were used, namely: Arts and Humanities, Entertainment, Computers and the Internet, Health, and Business and Economy. A total of 280 web sites were collected from Yahoo categories and classified into these categories. An additional 140 web sites were collected for use

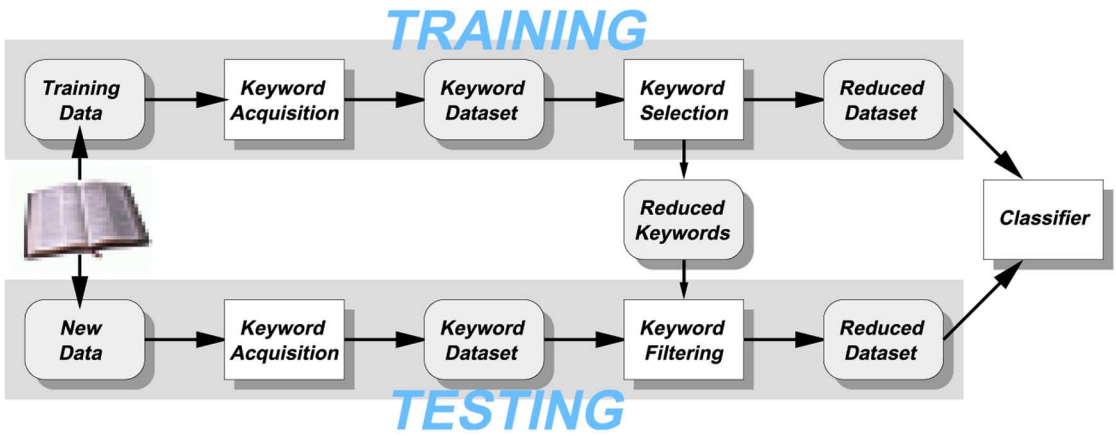

Fig. 6. Modular decomposition of the classification system.

TABLE X
PERFORMANCE: TRAINING DATA (USING VSM)

Method	Attributes	Average Precision	Average Error
Original	2557	-	-
Crisp	29	73.7%	23.8%
Fuzzy	23	78.1%	16.7%

TABLE XI
PERFORMANCE: UNSEEN DATA

Method	% Original Attributes	Classifier	Average Precision	Average Error
Crisp	1.13	BIM	23.3%	88.3%
		VSM	45.2%	49.7%
Fuzzy	0.90	BIM	16.7%	80.0%
		VSM	74.9%	35.9%

as test data. From this collection of data, the keywords, weights and corresponding classifications were collated into training and test datasets, containing 2557 attributes.

For the task of classification, two simple classifiers were used: the Boolean inexact model (BIM) [30] and the vector space model (VSM) [29]. More efficient and effective classifiers can be employed for this, but for simplicity only these conventional classifiers are adopted here to show the power of attribute reduction. Better classifiers are expected to produce more accurate results, though not necessarily enhance the comparisons between classifiers that use reduced or unreduced datasets.

A. Results

For this set of experiments, FRFS is compared with the standard crisp RSAR approach. As the unreduced training dataset exhibits high dimensionality (2557 attributes), it is too large to evaluate. This motivates the use of feature selection methods to reduce dimensionality to a more manageable size.

Using RSAR, the original dataset was reduced to 29 attributes (1.13% of the full attribute set). However, using FRFS the number of selected attributes was only 23 (0.90% of the full attribute set). It is interesting to note that the subsets discovered by FRFS and RSAR share four attributes in common. With such a large reduction in attributes, it must be shown that classification accuracy does not suffer in the FRFS-reduced system.

To see the effect of dimensionality reduction on classification accuracy, the system was tested on the original training data first and the results are summarised in Table X. The results are averaged over all the classification categories. Clearly, FRFS exhibits better precision and error rates. Note that this performance was achieved using *fewer* attributes than the crisp RSAR approach.

Table XI contains the results for experimentation on 140 previously unseen web sites. For the crisp case, the average precision is rather low and the average error is high. With FRFS, there is a significant improvement in both the precision and classification error.

It must be pointed out here that although the testing accuracy is rather low, this is largely to do with the poor performance of the simple classifiers used. The fact that VSM-based results are much better than those using BIM-based classifiers shows that when a more accurate classification system is employed, the accuracy can be considerably improved with the involvement of the same attributes. Nevertheless, the purpose of the present experimental studies is to compare the performance of the two attribute reduction techniques, based on the common use of any given classifier. Thus, only the relative accuracies are important. Also, it is worth noting that the classifications were checked automatically. Many websites can be classified to more than one category, however only the designated category is considered to be correct here.

FRFS requires a reasonable fuzzification of the input data, whilst the fuzzy sets are herein generated by simple statistical analysis of the dataset with no attempt made at optimizing these sets. A fine-tuned fuzzification will certainly improve the performance of FRFS-based systems [21].

VII. APPLICATION TO COMPLEX SYSTEMS MONITORING

The ever-increasing demand for dependable, trustworthy intelligent diagnostic and monitoring systems, as well as knowledge-based systems in general, has focused much of the attention of researchers on the knowledge-acquisition bottleneck. The task of gathering information and extracting general knowledge from it is known to be the most difficult part of creating a knowledge-based system. Complex application problems, such

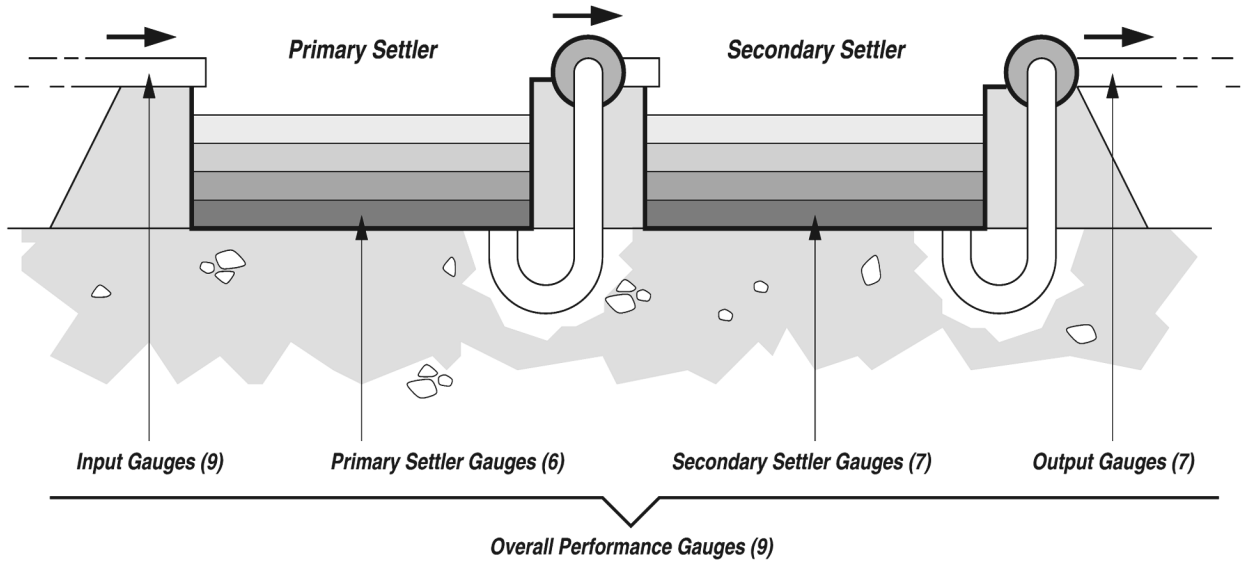

Fig. 7. Water treatment plant, with number of measurements shown at different points in the system.

as reliable monitoring and diagnosis of industrial plants, are likely to present large numbers of attributes, many of which will be redundant for the task at hand [24], [32]. This greatly hinders the performance of rule induction algorithms (RIAs). Additionally, inaccurate and/or uncertain values cannot be ruled out. Such applications typically require convincing explanations about the inference performed, therefore a method to allow automated generation of knowledge models of clear semantics is highly desirable.

In order to speed up the rule induction task and reduce rule complexity, a preprocessing step is required. This is particularly important for tasks where learned rulesets need regular updating to reflect the changes in the description of domain attributes. This step reduces the dimensionality of potentially very large attribute sets while minimising the loss of information needed for rule induction. It has an advantageous side-effect in that it removes redundancy from the historical data. This also helps simplify the design and implementation of the actual pattern classifier itself, by determining what attributes should be made available to the system. In addition, the reduced input dimensionality increases the processing speed of the classifier, leading to better response times. Most significant, however, is the fact that fuzzy-rough attribute selection preserves the semantics of the surviving attributes after removing any redundant ones. This is essential in satisfying the requirement of user readability of the generated knowledge model, as well as ensuring the understandability of the pattern classification process.

A. The Application

In order to evaluate further the utility of the FRFS approach and to illustrate its domain-independence, a challenging test dataset was chosen, namely the Water Treatment Plant Database [4] (in addition to the experimental evaluation carried out in the last section).

1) Problem Case: The dataset itself is a set of historical data charted over 521 days, with 38 different input attributes measured daily. Each day is classified into one of thirteen categories depending on the operational status of the plant. However, these can be collapsed into just two or three categories (i.e., *Normal* and *Faulty*, or *OK*, *Good* and *Faulty*) for plant monitoring purposes as many classifications reflect similar performance. Because of the efficiency of the actual plant the measurements were taken from, all faults appear for short periods (usually single days) and are dealt with immediately. This does not allow for a lot of training examples of faults, which is a clear drawback if a monitoring system is to be produced. Note that this dataset has been utilised in many previous studies, including that reported in [33] (to illustrate the effectiveness of applying crisp RSAR as a preprocessing step to rule induction, where a different RIA is adopted from here).

The 38 conditional attributes account for the following five aspects of the water treatment plant's operation (see Fig. 7):

1) input to plant (nine attributes);
2) input to primary settler (six attributes);
3) input to secondary settler (seven attributes);
4) output from plant (seven attributes);
5) overall plant performance (nine attributes).

The original dataset was split into 75% training and 25% testing data, maintaining the proportion of classifications present. It is likely that not all of the 38 input attributes are required to determine the status of the plant, hence, the dimensionality reduction step. However, choosing the most informative attributes is a difficult task as there will be many dependencies between subsets of attributes. There is also a monetary cost involved in monitoring these inputs, so it is desirable to reduce this number.

This work follows the original approach for complex systems monitoring developed in [33]. The original monitoring system consisted of several modules as shown in Fig. 8. It is this modular structure that allows the new FRFS technique to replace the existing crisp method [34].

Originally, a precategorization step preceded attribute selection where attribute values were quantized. To reduce potential loss of information, the original use of just the dominant symbolic labels of the discretized fuzzy terms is now replaced by a fuzzification procedure. This leaves the underlying attribute

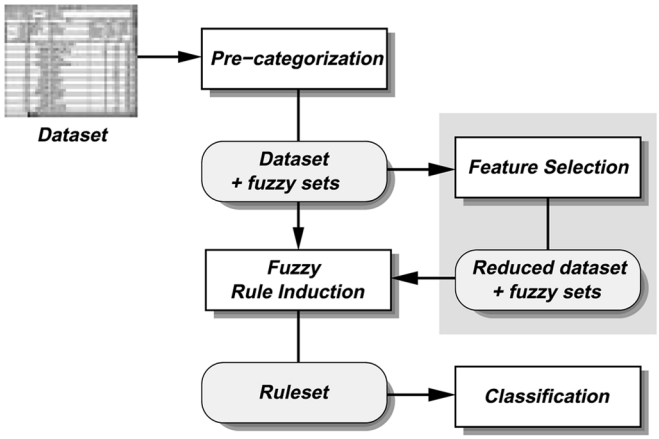

Fig. 8. Modular decomposition of the implemented system.

values unchanged but generates a series of fuzzy sets for each attribute. These sets are generated entirely from the data while exploiting the statistical data attached to the dataset (in keeping with the rough set ideology in that the dependence of learning upon information provided outside of the training dataset is minimized). This module may be replaced by alternative fuzzifiers, or expert-defined fuzzification if available. Based on these fuzzy sets and the original real-valued dataset, FRFS calculates a reduct and reduces the dataset accordingly. Finally, fuzzy rule induction is performed on the reduced dataset using the modelling algorithm developed in [6]. Note that this algorithm is not optimal, nor is the fuzzification. Yet the comparisons given later are fair due to their common background. Alternative fuzzy modelling techniques can be employed for this if available.

B. Experimental Results

The experiments were carried out over a tolerance range (with regard to the employment of the RIA). A suitable value for the threshold α must be chosen before rule induction can take place. However, the selection of α tends to be an application-specific task; a good choice for this threshold that provides a balance between a resultant ruleset's complexity and accuracy can be found by experiment. It should be noted here that due to the fuzzy rule induction method chosen, all approaches generate exactly the same number of rules (as the number of classes of interest), but the arities in different rulesets differ. This helps avoid a potential complexity factor in the comparative studies due to the need otherwise of considering the sizes of learned rulesets. Only the complexity in each learned rule needs to be examined,

C. Comparison With Unreduced Attributes

First of all, it is important to show that, at least, the use of attributes selected does not significantly reduce the classification accuracy as compared to the use of the full set of original attributes. For the 2-class problem, the fuzzy-rough set-based attribute selector returns 10 attributes out of the original 38.

Fig. 9 compares the classification accuracies of the reduced and unreduced datasets on both the training and testing data. As can be seen, the FRFS results are almost always better than the

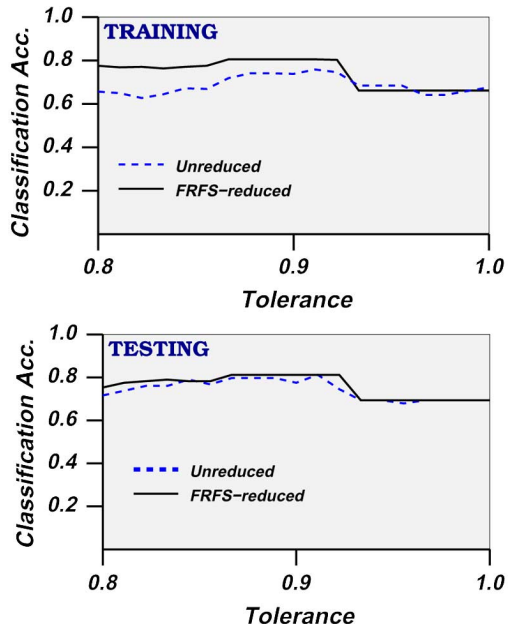

Fig. 9. Training and testing accuracies for the 2-class dataset over the tolerance range.

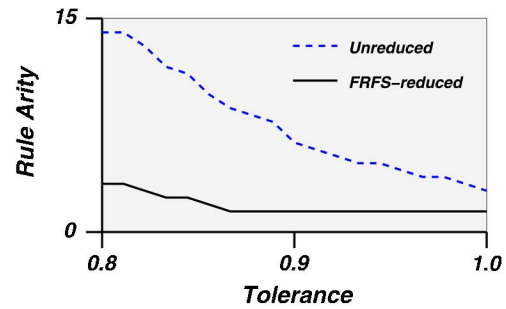

Fig. 10. Average rule arities for the 2-class dataset.

unreduced accuracies over the tolerance range. The best results for FRFS were obtained when α is in the range 0.86 to 0.90, producing a classification accuracy of 83.3% on the training set and 83.9% for the test data. Compare this with the optimum for the unreduced approach, which gave an accuracy of 78.5% for the training data and 83.9% for the test data.

By using the FRFS-based approach, rule complexity is greatly reduced. Fig. 10 charts the average rule complexity over the tolerance range for the two approaches. Over the range of α values, FRFS produces significantly less complex rules while having a higher resultant classification accuracy. The average rule arity of the FRFS optimum is 1.5 ($\alpha \in (0.86, 0.9)$) which is less than that of the unreduced optimum, 6.0.

The 3-class dataset is a more challenging problem, reflected in the overall lower classification accuracies produced. The fuzzy-rough method chooses 11 out of the original 38 attributes. The results of both approaches are presented in Fig. 11. Again, it can be seen that FRFS outperforms the unreduced approach on the whole. The best classification accuracy obtained for FRFS was 70.0% using the training data, 71.8% for the test data ($\alpha = 0.81$). For the unreduced approach, the best accuracy obtained was 64.4% using the training data, 64.1% for the test data ($\alpha = 0.88$).

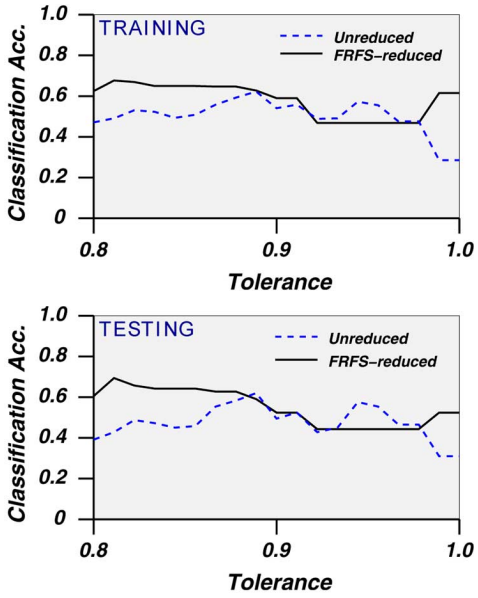

Fig. 11. Training and testing accuracies for the 3-class dataset over the tolerance range.

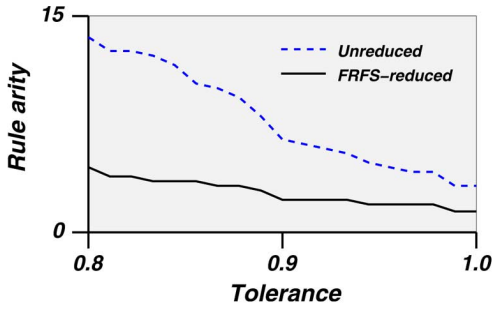

Fig. 12. Average rule arities for the 3-class dataset.

Fig. 12 compares the resulting rule complexity of the two approaches. It is evident that rules induced using FRFS as a preprocessor are simpler, with little loss in classification accuracy. In fact, the simple rules produced regularly outperform the more complex ones generated by the unreduced approach. The average rule arity for the FRFS-based method is 4.0 which is less than that of the unreduced method, 8.33.

These results show that FRFS is useful not only in removing redundant attribute measures but also in dealing with the noise associated with such measurements. The rules produced are reasonably short and understandable. However, when semantics-destroying dimensionality reduction techniques are applied, such readability is lost.

D. Comparison With Entropy-Based Attribute Selection

To support the study of the performance of FRFS for use as a preprocessor to rule induction, a conventional entropy-based technique is herein used for comparison. This approach utilizes the entropy heuristic employed by machine learning techniques such as C4.5 [28]. Those attributes that provide the most gain in information are selected. A summary of the results of this comparison can be seen in Table XII. Further related experimentation using C4.5 as the classification method can be found

TABLE XII
COMPARISON OF FRFS AND ENTROPY-BASED ATTRIBUTE SELECTION

Approach	No. of Classes	No. of Attributes	Training Accuracy	Testing Accuracy
FRFS	2	10	83.3%	83.9%
Entropy	2	13	80.7%	83.9%
FRFS	3	11	70.0%	71.8%
Entropy	3	14	70.0%	72.5%

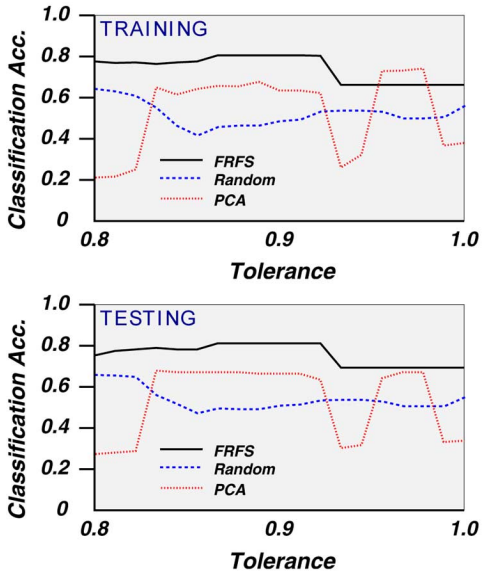

Fig. 13. Training and testing accuracies for the 2-class dataset: comparison with PCA and random-reduction methods.

in [17], where FRFS and entropy-based selection are compared with a novel ant colony optimization-based method.

For both the 2-class and 3-class datasets, FRFS selects three fewer attributes than the entropy-based method. FRFS has a higher training accuracy and the same testing accuracy for the 2-class data using less attributes. However, for the 3-class data, the entropy-based method produces a very slightly higher testing accuracy. Again, it should be noted that this is obtained with three additional attributes over the FRFS approach.

E. Comparison With PCA and Random Reduction

The previous comparisons ensured that little information loss is incurred due to FRFS. The question now is whether any other attribute sets of a dimensionality 10 (for the 2-class dataset) and 11 (for the 3-class dataset) would perform similarly. To avoid a biased answer to this, without resorting to exhaustive computation, 70 sets of random reducts were chosen of size 10 for the 2-class dataset, and a further 70 of size 11 for the 3-class dataset to see what classification results might be achieved. The classification accuracies for each tolerance value are averaged.

The effect of using a different dimensionality reduction technique, namely PCA, is also investigated. To ensure that the comparisons are fair, only the first 10 principal components are chosen for the 2-class dataset (likewise, the first 11 for the 3-class dataset). As PCA irreversibly destroys the underlying dataset semantics, the resulting rules are not human-comprehensible but may still provide useful automatic classifications of new data.

TABLE XIII
BEST INDIVIDUAL CLASSIFICATION ACCURACIES (2-CLASS DATASET) FOR FRFS, PCA, AND RANDOM APPROACHES

Approach	Training Accuracy	Testing Accuracy
FRFS	83.3%	83.9%
Random	66.4%	68.1%
PCA	76.7%	70.3%

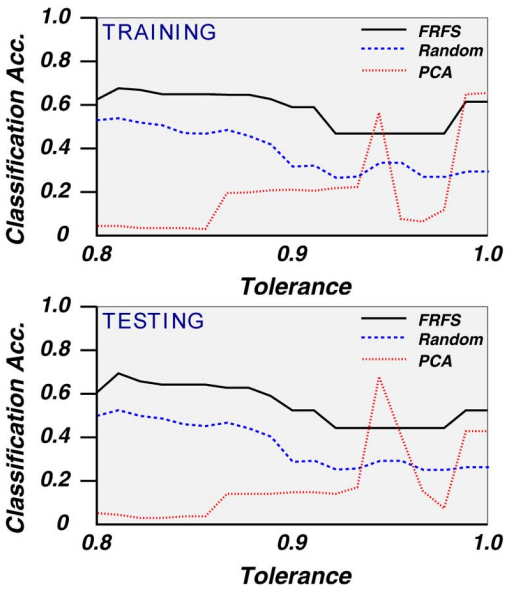

Fig. 14. Training and testing accuracies for the 3-class dataset: comparison with PCA and random-reduction methods.

The results of FRFS, PCA and random approaches can be seen in Fig. 13 for the 2-class dataset. On the whole, FRFS produces a higher classification accuracy than both PCA-based and random-based methods over the tolerance range. In fact, FRFS results in the highest individual classification accuracy for training and testing data (see Table XIII).

For the 3-class dataset, the results of FRFS, PCA and random selection are shown in Fig. 14. The individual best accuracies can be seen in Table XIV. Again, FRFS produces the highest classification accuracy (71.8%), and is almost always the best over the tolerance range. Although PCA produces a comparatively reasonable accuracy of 70.2%, this is at the expense of incomprehensible rules.

F. Alternative Fuzzy Rule Inducer

As stated previously, many fuzzy rule induction algorithms exist and can be used to replace the RIA adopted in the present monitoring system. Here, an example is given using Lozowski's algorithm as presented in [20]. This method extracts linguistically expressed fuzzy rules from real-valued attributes as with the subsethood-based RIA. Provided with training data, it induces approximate relationships between the characteristics of the conditional attributes and their underlying classes. However, as with many RIAs, this algorithm exhibits high computational complexity due to its generate-and-test nature. The effects of this become evident where high dimensional data needs to be processed. Indeed, for this particular domain, attribute selection is essential as running the RIA on all conditional attributes would be computationally prohibitive.

TABLE XIV
BEST RESULTANT CLASSIFICATION ACCURACIES (3-CLASS DATASET) FOR FRFS, PCA, AND RANDOM APPROACHES

Approach	Training Accuracy	Testing Accuracy
FRFS	70.0%	71.8%
Random	55.7%	54.3%
PCA	67.7%	70.2%

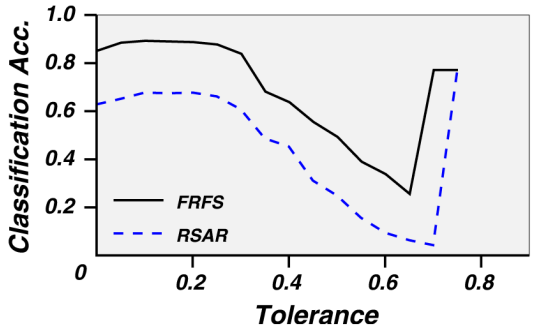

Fig. 15. Classification accuracies for the 2-class dataset.

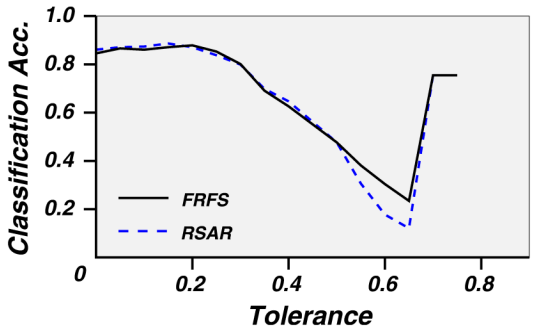

Fig. 16. Classification accuracies for the 3-class dataset.

TABLE XV
EXTENT OF DIMENSIONALITY REDUCTION

Method	Attributes 2-class	Attributes 3-class
FRFS	10	11
RSAR	11	11

The results presented here compare the use of fuzzy-rough set based attribute selection with the crisp rough set-based method. For RSAR, the data is discretized using the supplied fuzzy sets and reduction performed on the resulting dataset. The experiments were carried out over a tolerance range, required by the fuzzy RIA. This is a different threshold from those required in the subsethood-based approach. The tolerance here indicates the minimal confidence gap in the decision between a candidate rule and other competing contradictory rules.

As can be seen from Table XV, FRFS selects fewer attributes than the crisp method for the 2-class dataset and results in a higher classification accuracy over the entire tolerance range (Fig. 15). Both results show that there is a lot of redundancy in the dataset which may be removed with little loss in classification accuracy.

TABLE XVI
ATTRIBUTE EVALUATION FOR $x > 0.5, y = \sqrt{x}, z = x^2$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.5257	0.31758	0.997	1	99.5	200
y	0.5296	0.24586	0.997	1	99.5	200
z	0.5809	0.32121	0.997	1	99.5	200
3	0.0	-0.00276	0	0	55.5	0
4	0.0	0.00148	0	0	47.5	0
5	0.0	-0.00268	0	0	44.5	0
6	0.0	-0.00221	0	0	58.5	0
7	0.0	-0.01002	0	0	52.5	0
8	0.0	-0.00649	0	0	57.5	0
9	0.0	0.00889	0	0	49	0
10	0.0	-0.00222	0	0	53	0
11	0.0	0.00182	0	0	59.5	0
12	0.0	0.00144	0	0	42.5	0
13	0.0	0.00475	0	0	50	0
14	0.0	0.01006	0	0	65.5	0
15	0.0	0.00613	0	0	55.5	0
16	0.0	0.00488	0	0	47	0
17	0.0	0.00563	0	0	56.5	0
18	0.0	0.01427	0	0	50	0
19	0.0	-0.00467	0	0	53.5	0
20	0.0	-0.01785	0	0	54.5	0
21	0.0	0.00327	0	0	50	0
22	0.0	0.0035	0	0	48.5	0
23	0.0	0.01339	0	0	51.5	0
24	0.0	0.00464	0	0	49.5	0
25	0.0	0.01334	0	0	59.0	0
26	0.0	0.01715	0	0	48.5	0
27	0.0	-0.01742	0	0	49	0
28	0.0	0.00685	0	0	60.5	0
29	0.0	-0.00206	0	0	53.5	0
30	0.0	0.00164	0	0	51.5	0
31	0.0	0.00171	0	0	49	0
32	0.0	-0.00325	0	0	51	0

TABLE XVII
ATTRIBUTE EVALUATION FOR $(x + y)^2 > 0.25$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.2330	0.1862	0.2328	0.1579	86.75	128.47
y	0.2597	0.1537	0.1687	0.169	87.75	71.97
2	0.0	0.0132	0	0	84.5	0
3	0.0	0.0307	0	0	85.25	0
4	0.0	0.032	0	0	86	0
5	0.0	0.0112	0	0	85.5	0
6	0.0	0.0127	0	0	86	0
7	0.0	0.0248	0	0	86	0
8	0.0	0.0219	0	0	86	0
9	0.0	0.0364	0	0	85.5	0
10	0.012	0.0345	0.0344	0.0576	86.5	23.64
11	0.0	0.018	0	0	85.5	0
12	0.0	0.0246	0	0	85.75	0
13	0.0	0.0312	0	0	86	0
14	0.0	0.0182	0	0	86	0
15	0.0	0.0216	0	0	86	0
16	0.0	0.0245	0	0	86	0
17	0.0	0.0188	0	0	85.25	0
18	0.0	0.0145	0	0	85.5	0
19	0.0	0.0292	0	0	86	0
20	0.0	0.0132	0	0	86	0
21	0.0	0.0148	0	0	84.5	0
22	0.0	0.0116	0	0	86	0
23	0.0	0.0248	0	0	86	0
24	0.0	0.019	0	0	86	0
25	0.0	0.029	0	0	85.25	0
26	0.0	0.0222	0	0	86	0
27	0.0	0.0292	0	0	85.75	0
28	0.0	0.0307	0	0	84.75	0
29	0.0	0.0255	0	0	86	0
30	0.0	0.0163	0	0	86	0
31	0.0	0.0221	0	0	84.5	0

TABLE XVIII
ATTRIBUTE EVALUATION FOR $(x + y)^2 > 0.5$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.209	0.140067	0.241	0.156	79	119.56
y	0.2456	0.151114	0.248	0.165	78.25	122.34
2	0.0	0.00845	0	0	76	0
3	0.0	0.009063	0	0	73.75	0
4	0.0	0.005004	0	0	70.25	0
5	0.0	0.013202	0	0	74.75	0
6	0.0	0.011766	0	0	72.25	0
7	0.0	0.029141	0	0	73.5	0
8	0.0	0.007746	0	0	74.25	0
9	0.0	0.007245	0	0	73.5	0
10	0.0	0.018969	0	0	76.25	0
11	0.0	0.008741	0	0	75.5	0
12	0.0	0.012712	0	0	72.5	0
13	0.0	0.009962	0	0	72.25	0
14	0.0	-0.000115	0	0	75	0
15	0.0	0.003541	0	0	73.5	0
16	0.0	0.012629	0	0	75	0
17	0.0	0.019661	0	0	73.75	0
18	0.0	0.013886	0	0	76	0
19	0.0	0.011437	0	0	73.25	0
20	0.0	0.008366	0	0	74.25	0
21	0.0	0.017771	0	0	72.25	0
22	0.0	0.00363	0	0	74.5	0
23	0.0	0.013811	0	0	75.5	0
24	0.0	0.01756	0	0	74.5	0
25	0.0	0.003648	0	0	73.5	0
26	0.0	0.013574	0	0	72.75	0
27	0.0	0.009583	0	0	73.75	0
28	0.0	-0.000367	0	0	75.25	0
29	0.0	-0.000397	0	0	75.25	0
30	0.0	0.011544	0	0	76.25	0
31	0.0	0.007605	0	0	74.75	0

For the 3-class dataset the approaches perform similarly, with the FRFS method generally outperforming the other two, using the same number of attributes (but not *identical* attributes). The classification results can be seen in Fig. 16.

VIII. CONCLUSION

This paper has been concerned with the development of fuzzy-rough attribute selection, combatting the problems of noisy and real-valued data, as well as handling mixtures of nominal and continuous valued attributes. FRFS achieves this by the use of fuzzy-rough sets, and the new measure of attribute significance: the fuzzy-rough degree of dependency. A particular issue for attribute selectors is the problem of real-valued decision attributes. FRFS can deal with this whereas many AS techniques cannot.

The new fuzzy-rough metric was experimentally evaluated against other leading metrics for use in attribute ranking. The results confirmed that the fuzzy-rough measure performs comparably to these metrics, and better than them in several cases.

The dimensionality of the datasets involved in text categorisation are of the order of thousands to tens of thousands. FRFS was used to tackle this potentially restrictive amount of data successfully within a web page categorisation system. In fact, the extent of data reduction was several orders of magnitude, making the classification task manageable. The fuzzy-rough technique was also applied to complex systems monitoring to show how not only rule clarity can be significantly improved with attribute selection, but also that the reduced knowledge base can achieve competitive results in terms of monitoring accuracy. The fuzzy-rough method was shown to perform very well against other attribute selector methods for this task.

Through this series of investigations and experiments, the potential utility of the fuzzy-rough method for attribute selection

TABLE XIX
ATTRIBUTE EVALUATION FOR $(x + y)^3 < 0.125$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.2445	0.1486	0.134	0.134	87.75	57.46
y	0.2441	0.1659	0.159	0.164	87.25	73.39
2	0.0	0.0229	0	0	88.5	0
3	0.0	0.0232	0	0	89	0
4	0.0	0.0322	0	0	88.25	0
5	0.0	0.0301	0	0	89	0
6	0.0	0.0252	0	0	89	0
7	0.0	0.0203	0	0	89	0
8	0.0	0.0341	0	0	89	0
9	0.0	0.0289	0	0	89	0
10	0.0	0.0339	0	0	88.5	0
11	0.0	0.0313	0	0	89	0
12	0.0	0.0287	0	0	89	0
13	0.0	0.0545	0	0	89	0
14	0.0	0.0458	0	0	89	0
15	0.0	0.0378	0	0	89	0
16	0.0	0.0289	0	0	89	0
17	0.0	0.0332	0	0	89	0
18	0.0	0.0306	0	0	89	0
19	0.0	0.0397	0	0	88.25	0
20	0.0	0.0247	0	0	89	0
21	0.0	0.0163	0	0	89	0
22	0.0	0.033	0	0	89	0
23	0.0	0.0276	0	0	89	0
24	0.0	0.0189	0	0	88.75	0
25	0.0	0.0279	0	0	88.75	0
26	0.0	0.0252	0	0	88.75	0
27	0.0	0.0157	0	0	89	0
28	0.0	0.0304	0	0	89	0
29	0.0	0.0285	0	0	89	0
30	0.0	0.0315	0	0	88.75	0
31	0.0	0.029	0	0	89	0

TABLE XX
ATTRIBUTE EVALUATION FOR $x * y * z > 0.125$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.1057	0.0750547	0.169	0.123	64.25	73.65
y	0.0591	0.1079423	0.202	0.226	66.75	88.04
z	0.1062	0.0955878	0.202	0.16	67.5	84.28
3	0.0	0.003139	0	0	56.75	0
4	0.0	-0.0156922	0	0	60.75	0
5	0.0	0.0088234	0	0	58.5	0
6	0.0	-0.0076636	0	0	53.25	0
7	0.0	0.0050098	0	0	57.5	0
8	0.0	0.0006841	0	0	55.75	0
9	0.0	-0.0015287	0	0	54	0
10	0.0	0.0031223	0	0	53	0
11	0.0	0.0021915	0	0	57.75	0
12	0.0	0.002726	0	0	61.75	0
13	0.0	0.0108794	0	0	57.75	0
14	0.0	0.0008456	0	0	59.25	0
15	0.0	-0.000293	0	0	60	0
16	0.0	-0.001822	0	0	57.5	0
17	0.0	0.0019899	0	0	61.75	0
18	0.0	0.0090028	0	0	57.5	0
19	0.0	0.0043929	0	0	60.25	0
20	0.0	0.0006062	0	0	53.75	0
21	0.0	-0.0075626	0	0	53.75	0
22	0.0	0.0185202	0	0	57.0	0
23	0.0	-0.0056034	0	0	59.25	0
24	0.0	0.0116144	0	0	57.75	0
25	0.0	0.0001139	0	0	55.75	0
26	0.0	-0.0010561	0	0	56.25	0
27	0.0	0.0002921	0	0	54.5	0
28	0.0	0.0062014	0	0	51.75	0
29	0.0	-0.0092218	0	0	59.25	0
30	0.0	0.0000525	0	0	61.75	0
31	0.0	-0.001146	0	0	57.0	0
32	0.0	-0.0059597	0	0	57.0	0

has been demonstrated. However, there are many other areas that benefit from a data reduction step. It would be highly beneficial to investigate how FRFS may be applied to other domains

TABLE XXI
ATTRIBUTE EVALUATION FOR $x * y * z^2 > 0.125$

Attribute	FR	Re	IG	GR	1R	χ^2
x	0.1511	0.098	0.1451	0.0947	76.5	65.43
y	0.1101	0.05571	0.0909	0.108	78	35.36
z	0.2445	0.14736	0.2266	0.2271	79.75	93.81
3	0.0	0.00725	0	0	77.5	0
4	0.0	0.00652	0	0	78.5	0
5	0.0	0.01793	0	0	77.75	0
6	0.0	0.00716	0	0	78	0
7	0.0	0.02053	0	0	76.5	0
8	0.0	0.00339	0	0	78.25	0
9	0.0	0.01114	0	0	77	0
10	0.0	0.00409	0	0	77.75	0
11	0.0	0.01595	0	0	77.75	0
12	0.0	0.0164	0	0	77.75	0
13	0.0	0.01224	0	0	78.5	0
14	0.0	0.0017	0	0	75.5	0
15	0.0	0.00735	0	0	78.75	0
16	0.0	0.00575	0	0	78	0
17	0.0	0.01831	0	0	78.25	0
18	0.0	0.00508	0	0	76	0
19	0.0	0.01943	0	0	79	0
20	0.0	0.00929	0	0	78.5	0
21	0.0	0.00493	0	0	77.75	0
22	0.0	0.00579	0	0	75.75	0
23	0.0	0.01252	0	0	76.25	0
24	0.0	0.01957	0	0	79	0
25	0.0	0.017	0	0	78.25	0
26	0.0	0.01175	0	0	76.5	0
27	0.0	0.01055	0	0	76.5	0
28	0.0	0.01405	0	0	78	0
29	0.0	0.02123	0	0	77.75	0
30	0.0	0.00884	0	0	77.5	0
31	0.0	0.0127	0	0	77.75	0
32	0.0	0.00806	0	0	77.5	0

such as image recognition and gene expression analysis. For example, gene expression microarrays are a rapidly maturing technology that provide the opportunity to analyse the expression levels of thousands or tens of thousands of genes in a single experiment. As a result of the high dimensionality of this type of data, attribute selection must take place before any further processing can be carried out.

Part of the future work in fuzzy-rough feature selection would be to consider other alternatives to fuzzy similarity. For example, the strong transitivity condition based on the minimum operator in the fuzzy similarity relation definition could be replaced with the Lukasiewicz triangular norm. This would result in a so-called likeness function, and may result in more flexibility when dealing with uncertainty.

REFERENCES

[1] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Proc. 9th Nat. Conf. Artif. Intell.*, 1991, pp. 547–552.
[2] "Rough sets and current trends in computing," in *Proc. 3rd Int. Conf.*, J. J. Alpigini, J. F. Peters, J. Skowronek, and N. Zhong, Eds., 2002.
[3] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton Univ. Press, 1961.
[4] C. L. Blake and C. J. Merz, *UCI Repository of Machine Learning Databases*. Irvine, CA: Univ. California, 1998 [Online]. Available: http://www.ics.uci.edu/~mlearn/
[5] M. A. Carreira-Perpinñán, "Continuous latent variable models for dimensionality reduction and sequential data reconstruction," Ph.D. dissertation, Univ. Sheffield, Sheffield, U.K., 2001.
[6] S. Chen, S. L. Lee, and C. Lee, "A new method for generating fuzzy rules from numerical data for handling classification problems," *Appl. Artif. Intell.*, vol. 15, no. 7, pp. 645–664, 2001.
[7] A. Chouchoulas and Q. Shen, "Rough set-aided keyword reduction for text categorisation," *Appl. Artif. Intell.*, vol. 15, no. 9, pp. 843–873, 2001.

[8] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 3, 1997.
[9] D. Dubois and H. Prade, "Putting rough sets and fuzzy sets together," *Intell. Decision Support*, pp. 203–232, 1992.
[10] I. Düntsch and G. Gediga, *Rough Set Data Analysis: A Road to Non-Invasive Knowledge Discovery*. Bangor, ME: Methodos, 2000.
[11] B. S. Everitt, "An introduction to latent variable models," in *Monographs on Statistics and Applied Probability*. London, U.K.: Chapman & Hall, 1984.
[12] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Mach. Learn.*, vol. 11, pp. 63–90, 1993.
[13] E. Hunt, J. Martin, and P. Stone, *Experiments in Induction*. New York: Academic, 1966.
[14] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proc. 7th IEEE Int. Conf. Tools Artif. Intell.*, 1995, pp. 336–391.
[15] R. Jensen and Q. Shen, "Fuzzy-rough attribute reduction with application to web categorization," *Fuzzy Sets Syst.*, vol. 141, no. 3, pp. 469–485, 2004.
[16] ——, "Semantics-preserving dimensionality reduction: Rough and fuzzy-rough based approaches," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1457–1471, Dec. 2004.
[17] ——, "Fuzzy-rough data reduction with ant colony optimization," *Fuzzy Sets Syst.*, vol. 149, no. 1, pp. 5–20, 2005.
[18] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. 9th Nat. Conf. Artif. Intell.*, 1992, pp. 129–134.
[19] P. Langley, "Selection of relevant features in machine learning," in *Proc. AAAI Fall Symp. Relevance*, 1994, pp. 1–5.
[20] A. Lozowski, T. J. Cholewo, and J. M. Zurada, "Crisp rule extraction from perceptron network classifiers," in *Proc. Int. Conf. Neural Networks, volume of Plenary, Panel and Special Sessions*, 1996, pp. 94–99.
[21] J. G. Marin-Blázquez and Q. Shen, "From approximative to descriptive fuzzy classifiers," *IEEE Trans. Fuzzy Syst.*, vol. 10, no. 4, pp. 484–497, Aug. 2002.
[22] *Rough-Fuzzy Hybridization: A New Trend in Decision Making*. S. K. Pal and A. Skowron, Eds. New York: Springer-Verlag, 1999.
[23] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Norwell, MA: Kluwer, 1991.
[24] W. Pedrycz and G. Vukovich, "Feature analysis through information granulation," *Pattern Recogn.*, vol. 35, no. 4, pp. 825–834, 2002.
[25] "Rough set methods and applications: New developments in knowledge discovery in information systems," in *Studies in Fuzziness and Soft Computing*. L. Polkowski, T. Y. Lin, and S. Tsumoto, Eds. Berlin, Germany: Physica-Verlag, 2000, vol. 56.
[26] L. Polkowski, "Rough Sets: Mathematical Foundations," in *Advances in Soft Computing*. Berlin, Germany: Physica-Verlag, 2002.
[27] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, pp. 81–106, 1986.
[28] J. R. Quinlan, "C4.5: Programs for machine learning," in *The Morgan Kaufmann Series in Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
[29] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.
[30] G. Salton, E. A. Fox, and H. Wu, "Extended Boolean information retrieval," *Commun. ACM*, vol. 26, no. 12, pp. 1022–1036, 1983.
[31] J. C. Schlimmer, "Efficiently inducing determinations—A complete and systematic search algorithm that uses optimal pruning," in *Proc. Int. Conf. Mach. Learn.*, 1993, pp. 284–290.
[32] M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recogn.*, vol. 35, no. 4, pp. 835–846, 2002.
[33] Q. Shen and A. Chouchoulas, "A fuzzy-rough approach for generating classification rules," *Pattern Recogn.*, vol. 35, no. 11, pp. 341–354, 2002.
[34] Q. Shen and R. Jensen, "Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring," *Pattern Recogn.*, vol. 37, no. 7, pp. 1351–1363, 2004.
[35] W. Siedlecki and J. Sklansky, "On automatic feature selection," *Int. J. Pattern Recogn. Artif. Intell.*, vol. 2, pp. 197–220, 1988.
[36] A. Skowron and J. Stepaniuk, "Tolerance approximation spaces," *Fundamenta Informaticae*, vol. 27, no. 2, pp. 245–253, 1996.
[37] *Intelligent Decision Support*R. Slowinski, Ed. Norwell, MA: Kluwer Academic Publishers, 1992.
[38] R. Slowinski and D. Vanderpooten, "Similarity relation as a basis for rough approximations," in *Advances in Machine Intelligence and Soft Computing.* , P. Wang, Ed. Durham, NC: Duke Univ. Press, 1997, vol. IV, pp. 17–33.
[39] J. Stefanowski and A. Tsoukiàs, "Valued tolerance and decision rules," *Rough Sets Current Trends Comput.*, pp. 212–219, 2000.
[40] R. W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recogn. Lett.*, vol. 24, no. 6, pp. 833–849, 2003.
[41] H. Thiele, Fuzzy rough sets versus rough fuzzy sets—An interpretation and a comparative study using concepts of modal logics Univ. of Dortmund, Dortmund, Germany, Tech. Rep. CI-30/98, 1998.
[42] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools With Java Implementations*. San Mateo, CA: Morgan Kaufmann, 2000.
[43] *Yahoo*, [Online]. Available: www.yahoo.com
[44] Y. Y. Yao, "A comparative study of fuzzy sets and rough sets," *Inform. Sci.*, vol. 109, pp. 21–47, 1998.
[45] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. 14th Int. Conf. Mach. Learn.*, 1997, pp. 412–420.
[46] L. A. Zadeh, "Fuzzy sets," *Inform. Control*, vol. 8, pp. 338–353, 1965.
[47] ——, "The concept of a linguistic variable and its application to approximate reasoning," *Inform. Sci.*, vol. 8, pp. 199–249, 1975, 301–357; vol. 9: 43–80.

Richard Jensen received the B.Sc. degree in computer science from Lancaster University, Lancaster, U.K., and the M.Sc. and Ph.D. degrees in artificial intelligence from the University of Edinburgh, Edinburgh, U.K., in 1999, 2000, and 2004, respectively.

He is a Postdoctoral Research Fellow with the Department of Computer Science at the University of Wales, Aberystwyth, U.K., 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 approximately 20 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 Defense 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, U.K. His research interests include fuzzy and imprecise modeling, model-based inference, pattern recognition, and knowledge refinement and reuse. He has published around 170 peer-refereed papers in academic journals and conferences on topics within artificial intelligence and related areas.

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 *Fuzzy Sets and Systems*.