Finding Fuzzy-Rough Reducts with Fuzzy Entropy

Neil Mac Parthaláin, Richard Jensen, and Qiang Shen

Abstract— **Dataset dimensionality is undoubtedly the single most significant obstacle which exasperates any attempt to apply effective computational intelligence techniques to problem domains. In order to address this problem a technique which reduces dimensionality is employed prior to the application of any classification learning. Such feature selection (FS) techniques attempt to select a subset of the original features of a dataset which are rich in the most useful information. The benefits can include improved data visualisation and transparency, a reduction in training and utilisation times and potentially, improved prediction performance. Methods based on fuzzy-rough set theory have demonstrated this with much success. Such methods have employed the dependency function which is based on the information contained in the lower approximation as an evaluation step in the FS process. This paper presents three novel feature selection techniques employing fuzzy entropy to locate fuzzy-rough reducts. This approach is compared with two other fuzzy-rough feature selection approaches which utilise other measures for the selection of subsets.**

I. Introduction

When data is collected or recorded, very often every single aspect of the domain which is being examined may be considered such that complete representation can be achieved, and also to ensure that no potentially useful information is lost. The disadvantage associated with recording such large numbers of domain attributes however means that any attempt to use machine learning tools to extract knowledge, results in very poor performance. Feature Selection (FS) [5] is a process which attempts to select features which are information-rich but also retain the original meaning of the features following reduction. It is unsurprising therefore, that feature selection has been applied to problems which have very large dimensionality (>10,000) [3]. Problems of such scale are usually outside the scope of most learning algorithms, and in the few instances where they are not, the learning algorithm will often find patterns that are spurious and invalid.

As mentioned previously, it may be expected that the inclusion of an increasing number of features should also improve the likelihood of the ability to distinguish between classes. This may not be the case however if the training data size does not also demonstrate a simultaneous significant increase with the addition of each feature. Most learning approaches utilise a reduction step to overcome such problems when dealing with large dimensionality. An efficient and effective method to achieve this therefore is clearly desirable.

Rough set theory (RST) [13] offers an alternative, and formal methodology that can be employed to reduce the dimensionality of datasets, as a preprocessing step to assist knowledge discovery methods for learning from data. It helps to select the most valuable features in a dataset, and does this without transforming the data, whilst at the same time attempting to minimise information loss during the selection process. Computationally, the approach is very efficient, and relies on simple set operations, which in-turn makes it suitable as a preprocessor for techniques that are significantly more complex. Unlike statistical correlation-reduction approaches [7], RST requires no human input or intervention. Most importantly however, it retains the underlying semantics of the data, which results in models that are more transparent to human scrutiny. The primary disadvantage associated with the RST approach lies in its inability to deal with real-valued data, and a number of extensions to the basic rough set model have been proposed in an attempt to address this shortcoming, e.g. [18]. These extensions whilst offering more flexibility, rely on a threshold value or other information which is non-data derived. This obviously is a departure from the RST tenet of using only the information contained in the data.

Other approaches focus on hybridizing RST with other techniques such that one technique complements the other. One such approach is the combination of RST with fuzzy set theory to create fuzzy-rough sets [9], [17]. Fuzzy-rough feature selection (FRFS) provides a means by which discrete or real-valued noisy data (or a mixture of both) can be effectively reduced without the need for user-supplied information. Additionally, this technique can be applied to data with continuous or nominal decision attributes, and as such can be applied to regression as well as classification. This paper proposes three new measures based on fuzzy entropy in order to locate small, yet information-rich, fuzzy-rough feature subsets.

This paper is structured as follows. The theoretical background is given in section II, providing the necessary details for crisp rough set theory and fuzzy-rough concepts. In the third section, the new developments for fuzzy-rough feature selection incorporating measures of fuzzy entropy are presented. Some initial experimentation is provided in section IV. The paper is concluded in section V.

II. Background

Although the principal focus of this paper lies in the use of the various fuzzy entropy based evaluation measures of fuzzy-rough sets for FS, an in-depth view of both the RST and fuzzy-rough methodologies is necessary in order to demonstrate the motivation for the investigation of the fuzzy entropy as an evaluation metric. It is perhaps worth noting at this point that this paper does not introduce a new feature

Neil Mac Parthaláin (email: nsm03@aber.ac.uk), Richard Jensen (email: rkj@aber.ac.uk) and Qiang Shen (email: qqs@aber.ac.uk), are with the Department of Computer Science, Aberystwyth University, Aberystwyth, Wales, UK.

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

selection method but rather proposes a new set of evaluation metrics that are used to measure subset 'goodness'.

RST is an extension of conventional set theory which supports approximations in decision making. A rough set is the approximation of a vague concept by a pair of precise concepts which are known as upper and lower approximations. The lower approximation is a definition of the collection of the domain objects which are known with absolute certainty to belong to the concept of interest, whilst the upper approximation is the set of those objects which possibly belong to the concept of interest.

A. Rough Set Attribute Reduction (RSAR)

At the heart of the RSAR approach 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 non-empty finite set of attributes so that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that a can take. For any $P \subseteq \mathbb{A}$, there exists an associated equivalence relation $IND(P)$:

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

The partition generated by $IND(P)$ is denoted $\mathbb{U}/IND(P)$ and is calculated as follows:

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

where,

$$\mathbb{U}/IND(\{a\}) = \{\{x|a(x) = b,\ x \in \mathbb{U}\}|b \in V_a\} \quad (3)$$

and,

$$A \otimes B = \{X \cap Y \mid \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\} \quad (4)$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_p$. Let $X \subseteq \mathbb{U}$. X can be approximated using only the information contained in P by constructing the P-*lower* and P-*upper* approximations of X:

$$\underline{P}X = \{x|[x]_p \subseteq X\} \quad (5)$$

$$\overline{P}X = \{x|[x]_p \cap X \neq \emptyset\} \quad (6)$$

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

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \quad (7)$$

$$NEG_P(Q) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/Q} \overline{P}X \quad (8)$$

$$BND_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \overline{P}X - \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \quad (9)$$

By employing this definition of the positive region it is possible to calculate the rough set degree of dependency of a set of attributes Q on a set of attributes P. This can be achieved as follows: For P, $Q \subseteq A$, it can be said that Q depends on P in a degree k $(0 \leq k \leq 1)$, this is denoted $P \Rightarrow_k Q$ if:

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|\mathbb{U}|} \quad (10)$$

Where, $|\cdot|$ denotes the cardinality of the relevant set.

The reduction of attributes or selection of survival features can be achieved through the comparison of equivalence relations generated by sets of attributes. Attributes are removed such that the reduced set provides identical predictive capability of the decision feature or features as that of the original or unreduced set of features. A $reduct$ can be defined as a subset of minimal cardinality R_{min} of the conditional attribute set where $\gamma_R(\mathbb{D}) = \gamma_{\mathbb{C}}(\mathbb{D})$.

The QUICKREDUCT algorithm shown in Fig. 1 [3] searches for a minimal subset without exhaustively generating all possible subsets. The search begins with an empty subset, attributes which result in the greatest increase in the rough set dependency value are added iteratively. This process continues until the search produces its maximum possible dependency value for that dataset $(\gamma_{\mathbb{C}}(\mathbb{D}))$. Note that this type of hill-climbing search does not guarantee a minimal subset and may only discover a local minimum.

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

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

Fig. 1. The QUICKREDUCT algorithm

B. Fuzzy-Rough Feature Selection (FRFS)

Previous work on fuzzy-rough feature selection used a fuzzy partitioning of the input space [17] in order to determine fuzzy equivalence classes. Alternative definitions for the fuzzy lower and upper approximations can be found in [16], where a T-transitive fuzzy similarity relation is used to approximate a fuzzy concept X:

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x, y), \mu_X(y)) \quad (11)$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in \mathbb{U}} T(\mu_{R_P}(x,y), \mu_X(y)) \quad (12)$$

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

$$\mu_{R_P}(x,y) = \bigcap_{a \in P} \{\mu_{R_a}(x,y)\} \quad (13)$$

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

$$\mu_{R_a}(x,y) = 1 - \frac{|a(x) - a(y)|}{|a_{max} - a_{min}|} \quad (14)$$

$$\mu_{R_a}(x,y) = \exp(-\frac{(a(x) - a(y))^2}{2{\sigma_a}^2}) \quad (15)$$

$$\mu_{R_a}(x,y) = \max(\min(\frac{(a(y) - (a(x) - \sigma_a))}{(a(x) - (a(x) - \sigma_a))}, \frac{((a(x) + \sigma_a) - a(y))}{((a(x) + \sigma_a) - a(x))}, 0) \quad (16)$$

where ${\sigma_a}^2$ is the variance of feature a. As these relations do not necessarily display T-transitivity, the fuzzy transitive closure must be computed for each attribute [6]. The combination of feature relations in equation (13) has been shown to preserve T-transitivity [19].

1) Reduction: In a similar way to the original RSAR approach, the fuzzy positive region [10] can be defined as:

$$\mu_{POS_{R_P}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{R_P}X}(x) \quad (17)$$

The resulting degree of dependency is:

$$\gamma'_P(\mathbb{D}) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_{R_P}(\mathbb{D})}(x)}{|\mathbb{U}|} \quad (18)$$

A fuzzy-rough reduct R can be defined as a minimal subset of features that preserves the dependency degree of the entire dataset, i.e. $\gamma'_R(\mathbb{D}) = \gamma'_{\mathbb{C}}(\mathbb{D})$. Based on this, a fuzzy-rough QUICKREDUCT algorithm can be constructed that operates in the same way as Fig. 1, but uses equation (18) to gauge subset quality. In [10], it has been shown that the dependency function is monotonic and that fuzzy discernibility matrices may also be used to discover reducts.

Core features may be determined by considering the change in dependency of the full set of conditional features when individual attributes are removed:

$$Core(\mathbb{C}) = \{a \in \mathbb{C} | \gamma'_{\mathbb{C}-\{a\}}(Q) < \gamma'_{\mathbb{C}}(Q)\} \quad (19)$$

III. FUZZY ENTROPY FEATURE SELECTION

This section presents some new evaluation metrics for fuzzy-rough feature selection, based on the fuzzy entropy measure. These metrics are applied to the fuzzy-rough lower approximation and also to the fuzzy-rough boundary region.

A. Fuzzy Boundary Region-based FS

The lower approximation contains information regarding the extent of certainty of object membership to a given concept. However, the upper approximation contains information regarding the degree of uncertainty of objects and hence this information can be used to discriminate between subsets. For example, two subsets may result in the same lower approximation but one subset may produce a smaller upper approximation. This subset will be more useful as there is less uncertainty concerning objects within the boundary region (the difference between upper and lower approximations).

Following the original rough set approach, the fuzzy-rough boundary region for a concept X can be defined by:

$$\mu_{BND_{R_P}(X)}(x) = \mu_{\overline{R_P}X}(x) - \mu_{\underline{R_P}X}(x) \quad (20)$$

When the decision feature is real-valued the same fuzzy similarity measure is employed, resulting in the relation $R_{\mathbb{D}}$ with foresets $D_1, D_2, ..., D_n$. The fuzzy-rough boundary region then becomes:

$$\mu_{BND_{R_P}(D_j)}(x) = \frac{\mu_{\overline{R_P}D_j}(x) - \mu_{\underline{R_P}D_j}(x)}{|D_j|} \quad (21)$$

for decision foreset D_j, where $|D_j|$ stands for the cardinality of D_j.

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

$$U_P(X) = \frac{\sum_{x \in \mathbb{U}} \mu_{BND_{R_P}(X)}(x)}{|\mathbb{U}|} \quad (22)$$

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

$$\lambda_P(\mathbb{D}) = \frac{\sum_{X \in \mathbb{U}/\mathbb{D}} U_P(X)}{|\mathbb{U}/\mathbb{D}|} \quad (23)$$

When the decision feature is fuzzy, this becomes:

$$\lambda_P(\mathbb{D}) = \frac{\sum_{D_j \in R_{\mathbb{D}}} U_P(D_j)}{\sum_{D_n \in R_{\mathbb{D}}} (|D_n|)^{-1}} \quad (24)$$

Obviously, this degenerates to the previous definition when dealing with crisp decisions. A QUICKREDUCT-style algorithm can be constructed for locating fuzzy-rough reducts based on this measure. Instead of maximising the dependency degree, the task of the algorithm is to minimize the total uncertainty degree. When this reaches the minimum for the dataset, a fuzzy-rough reduct has been found.

Theorem 1: B-FRFS monotonicity. Suppose that $P \subseteq \mathbb{C}$, a is an arbitrary conditional feature that belongs to the dataset and Q is the set of decision features. Then $\lambda_{P\cup\{a\}}(Q) \leq \lambda_P(Q)$.

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

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

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

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

It is known from Theorem 1 in [9] that $\mu_{R_{P\cup\{a\}}}(x,y) \leq \mu_{R_P}(x,y)$, so $\mu_{\overline{R_{P\cup\{a\}}}X}(x) \leq \mu_{\overline{R_P}X}(x)$. As $\mu_{\underline{R_{P\cup\{a\}}}X}(x) \geq \mu_{\underline{R_P}X}(x)$, then $\mu_{BND_{R_{P\cup\{a\}}}(X)}(x) \leq \mu_{\overline{BND_{R_P}(X)}}(x)$. Thus, $U_{P\cup\{a\}}(Q) \leq U_P(Q)$ and therefore $\lambda_{P\cup\{a\}}(Q) \leq \lambda_P(Q)$. ■

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

TABLE I
EXAMPLE DATASET

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

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

For object 4, this is

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

For the remaining objects, this is:

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

Hence, the ID for concept {1,3,6} is:

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

For concept {2, 4, 5}, the ID is:

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

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

$$\begin{aligned}\lambda_a(Q) &= \frac{\sum_{X\in\mathbb{U}/Q} U_a(X)}{|\mathbb{U}/Q|} \\ &= \frac{0.899+0.899}{2} \\ &= 0.899\end{aligned} \tag{25}$$

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

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

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

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

The subset $\{b, c\}$ results in the minimal imprecision degree for the dataset, and the algorithm terminates. Interestingly, this is the same subset as that chosen by the fuzzy lower approximation-based method above.

B. Integration of Fuzzy Entropy

In the above method, the overall uncertainty is evaluated by averaging the uncertainty of all decision concepts. The ID for a concept is itself an average measure of the *belonging* of objects to the fuzzy boundary region. A more appropriate way of measuring the uncertainty within the boundary region of a concept X is to calculate the fuzzy entropy:

$$U'_P(X) = \sum_{x\in\mathbb{U}} -\frac{\mu_{BND_{R_P}(X)}(x)}{|BND_{R_P}(X)|} \log_2 \frac{\mu_{BND_{R_P}(X)}(x)}{|BND_{R_P}(X)|} \tag{26}$$

$$\lambda'_P(\mathbb{D}) = \frac{\sum_{D_j\in R_\mathbb{D}} U'_P(D_j)}{\sum_{D_n\in R_\mathbb{D}} (|D_n|)^{-1}} \tag{27}$$

This will be minimized when all fuzzy boundary regions are empty, hence $\lambda'_P(\mathbb{D}) = \lambda_P(\mathbb{D}) = 0$ and therefore P must be a fuzzy-rough reduct.

C. Fuzzy-Rough Reduction with Fuzzy Entropy

Fuzzy entropy itself can be used to find fuzzy-rough reducts [11]. A subset $P \subseteq \mathbb{C}$ induces a fuzzy similarity relation (R_P) with corresponding foresets $F_1, F_2, ..., F_n$. Similarly, the foresets induced by the (fuzzy) decision feature $\mathbb{D}$ are $D_1, D_2, ..., D_n$. The fuzzy entropy for a foreset F_i can be defined as:

$$H(F_i) = \sum_{D_j\in R_\mathbb{D}} \frac{-p(D_j|F_i)\log_2 p(D_j|F_i)}{|D_j|} \tag{28}$$

where $p(D_j|F_i)$ is the relative frequency of foreset F_i with respect to the decision D_j, and is defined:

$$p(D_j|F_i) = \frac{|D_j \cap F_i|}{|F_i|} \tag{29}$$

Based on these definitions, the fuzzy entropy for an attribute subset P can be defined as follows:

$$E(P) = \sum_{F_i \in R_P} \frac{|F_i|}{\sum_{Y_i \in R_P} |Y_i|} H(F_i) \tag{30}$$

This fuzzy entropy is monotonic and can be used to gauge the utility of feature subsets in a similar way to that of the fuzzy-rough measure. By dividing the entropy by $\log_2(\sum_{D_n \in R_{\mathbb{D}}} (|D_n|)^{-1})$, the measure will be normalized. This can be integrated into a QUICKREDUCT-style algorithm, employing a greedy hill-climbing approach. Again, as the measure monotonically decreases with addition of features, the search algorithm seeks to minimize this value in a manner similar to the boundary region minimization approach.

Theorem 2: E-FRFS reducts are fuzzy-rough reducts. Suppose that $P \subseteq \mathbb{C}$. If $E(P) = 0$ then P is a fuzzy-rough reduct.

Proof: Equation (17) can be rewritten as [9]:

$$\mu_{POS_{R_P}(\mathbb{D})}(x) = \sup_{D_j} \sup_{F_i} \min(\inf_{y \in \mathbb{U}} I(\mu_{F_i}(y), \mu_{D_j}(y)))$$

If P is a fuzzy-rough reduct, then it must be the case that $F_i \subseteq D_j$ or $F_i \cap D_j = \emptyset$ $\forall$F$_i, D_j$. If $F_i \subseteq D_j$, then $p(D_j|F_i) = 1$, and if $F_i \cap D_j = \emptyset$, then $p(D_j|F_i) = 0$ $\forall F_i, D_j$. Therefore each $H(F_i) = 0$, and $E(P) = 0$. ■

1) Example: Returning to the example dataset in Table I, the fuzzy entropy measure is used to determine fuzzy-rough reducts. The algorithm begins with an empty subset, and considers the addition of individual features. The attribute that results in the greatest decrease in fuzzy entropy will ultimately be added to the reduct candidate. For attribute a, the fuzzy entropy is calculated as follows ($A = \{a\}$):

$$E(A) = \sum_{F_i \in R_A} \frac{|F_i|}{\sum_{Y_i \in R_A} |Y_i|} H(F_i)$$

Each foreset F_i corresponds to one row in the matrix R_A:

F_1	1.0	1.0	0.699	0.0	0.0	0.0
F_2	1.0	1.0	0.699	0.0	0.0	0.0
F_3	0.699	0.699	1.0	0.0	0.0	0.0
F_4	0.0	0.0	0.0	1.0	0.699	0.699
F_5	0.0	0.0	0.0	0.699	1.0	1.0
F_6	0.0	0.0	0.0	0.699	1.0	1.0

Considering F_1, H(F_1) must be calculated:

$$H(F_1) = \sum_{D_j \in R_{\mathbb{D}}} \frac{-p(D_j|F_1)\log_2 p(D_j|F_1)}{|D_j|}$$

Each foreset D_j corresponds to one row in the matrix $R_{\mathbb{D}}$:

D_1	1.0	0.0	1.0	0.0	0.0	1.0
D_2	0.0	1.0	0.0	1.0	1.0	0.0
D_3	1.0	0.0	1.0	0.0	0.0	1.0
D_4	0.0	1.0	0.0	1.0	1.0	0.0
D_5	0.0	1.0	0.0	1.0	1.0	0.0
D_6	1.0	0.0	1.0	0.0	0.0	1.0

For D_1:

$$H(D_1) = \frac{-p(D_1|F_1)\log_2 p(D_1|F_1)}{|D_1|}$$

$$= \frac{-(1.699/2.699)\log_2(1.699/2.699)}{3.0}$$

Calculating this for each D_j produces:

$$H(F_1) = 0.140{+}0.177{+}0.140{+}0.177{+}0.177{+}0.140 = 0.951$$

The procedure is repeated for each remaining foreset:

$$H(F_2) = 0.951, H(F_3) = 0.871, H(F_4) = 0.871,$$
$$H(F_5) = 0.951, H(F_6) = 0.951$$

Hence, the fuzzy entropy is:

$$\begin{aligned} E(A) &= \sum_{F_i \in R_A} \frac{|F_i|}{\sum_{Y_i \in R_A} |Y_i|} H(F_i) \\ &= 0.926 = \text{E}(\{\text{a}\}) \end{aligned} \tag{32}$$

Repeating this process for the remaining attributes gives:

$E(\{b\})$ = 0.921
$E(\{c\})$ = 0.738

From this it can be seen that attribute c will cause the greatest decrease in fuzzy entropy. This attribute is chosen and added to the potential reduct, $R \leftarrow R \cup \{c\}$. The process iterates and the two fuzzy entropy values calculated are

$E(\{a, c\})$ = 0.669
$E(\{b, c\})$ = 0.0

Adding attribute b to the reduct candidate results in the minimum entropy for the data, and the search terminates, outputting the subset $\{b, c\}$. The dataset can now be reduced to only those attributes appearing in the reduct.

D. Fuzzy-Rough Reduction with Fuzzy Gain Ratio

The Information Gain (IG) [15] is the expected reduction in entropy resulting from partitioning the dataset objects according to a particular feature. For the fuzzy case this can be expressed as:

$$IG(P \cup \{a\}) = E(P) - E(P \cup \{a\}) \tag{33}$$

One limitation of the IG measure is that it favours features with many values. The Gain Ratio (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. Again, for the fuzzy case this can be expressed as:

TABLE II
REDUCT SIZE AND TIME TAKEN

Dataset	Objects	Features	Reduct size				
			E	B	L	BE	GR
Cleveland	297	14	10	9	9	10	10
Glass	214	10	9	9	10	10	9
Heart	270	14	9	8	8	8	9
Ionosphere	230	35	8	9	9	10	8
Olitos	120	26	6	6	6	6	6
Water 2	390	39	7	7	7	7	7
Water 3	390	39	7	7	7	7	7
Web	149	2557	23	20	21	20	18
Wine	178	14	6	6	6	6	6

$$SP(Q) = \sum_{F_i \in R_Q} \frac{|F_i|}{\sum_{Y_i \in R_Q} |Y_i|} \log_2 \frac{|F_i|}{\sum_{Y_i \in R_Q} |Y_i|} \quad (34)$$

The Gain Ratio is then defined as follows:

$$GR(P \cup \{a\}) = \frac{IG(P \cup \{a\})}{SP(P \cup \{a\})} \quad (35)$$

When this is minimized, $P \cup \{a\}$ is a fuzzy-rough reduct due to the monotonicity of the fuzzy entropy measure. This metric is applied in the same manner as described previously for the feature selection approach.

IV. EXPERIMENTATION

This section presents the initial experimental evaluation of the selection methods for the task of pattern classification, over nine benchmark datasets obtained from [12] with two classifier learners.

A. Experimental Setup

For the fuzzy-rough methods, the Łukasiewicz fuzzy connectives are used, with fuzzy similarity defined in (16). After feature selection, the datasets are reduced according to the discovered reducts. These reduced datasets are then classified using the relevant classifier learning method.

Two learning mechanisms were employed to create classifiers for the purpose of evaluating the resulting subsets from the feature selection phase: JRip [4] and PART [20], [21]. 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. PART generates rules by means of repeatedly creating partial decision trees from data. The algorithm adopts a divide-and-conquer strategy such that it removes instances covered by the current ruleset during processing. Essentially, a classification rule is created by building a pruned tree for the current set of instances; the leaf with the highest coverage is promoted to a rule.

B. Experimental Results

Table II compares the reduct size for fuzzy entropy-based FS (E), fuzzy boundary region-based FS (B), fuzzy lower approximation-based FS (L), fuzzy boundary/entropy FS (BE) and fuzzy gain ratio FS (GR). It can be seen that the new entropy-based fuzzy-rough methods find smaller subsets in general (B, BE, GR). The fuzzy boundary region-based method finds smaller or equally-sized subsets than the L. This is to be expected, as B includes fuzzy upper approximation information in addition to that of the fuzzy lower approximation. The entropy-based methods perform similarly, with the fuzzy gain ratio measure finding the smallest subsets in general. This demonstrates the utility of considering the split information when evaluating subset quality.

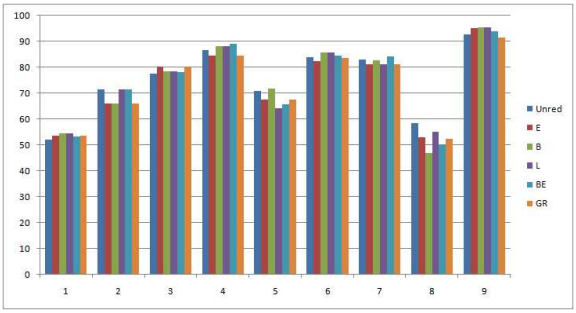

Fig. 2. Performance: JRip

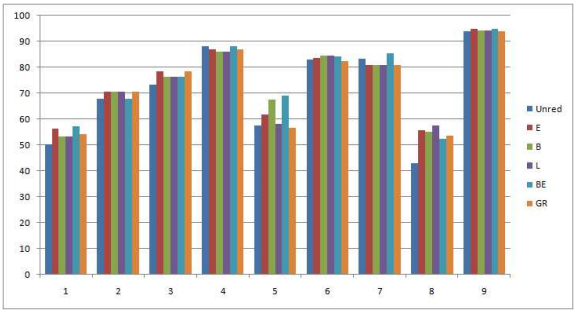

Fig. 3. Performance: PART

Table III shows the average classification accuracy as a percentage obtained using 10-fold cross validation. The classification accuracies are also presented in Figs. 2 and 3 for each of the nine datasets. The classification was initially performed on the unreduced dataset, followed by the reduced datasets which were obtained using the feature selection techniques. All techniques perform similarly, with both the boundary (B) and lower approximation (L) FS approaches showing the most consistent results for both classifier learners. It would appear that the GR approach also generally selects subsets at the expense of classification accuracy. The BE approach demonstrates that there is some useful

TABLE III
RESULTING CLASSIFICATION ACCURACIES (%)

Dataset	JRip						PART					
	Unred.	E	B	L	BE	GR	Unred.	E	B	L	BE	GR
Cleveland	52.19	53.53	54.55	54.55	53.20	53.53	50.17	56.22	53.20	53.20	57.23	56.22
Glass	71.50	65.89	65.89	71.50	71.50	65.89	67.76	70.56	70.56	67.76	67.76	70.56
Heart	77.41	80.37	78.52	78.52	78.15	80.37	73.33	78.51	76.30	76.30	76.30	78.51
Ionosphere	86.52	84.37	88.26	88.26	89.15	84.37	88.26	86.95	86.09	86.09	88.26	86.95
Olitos	70.83	67.50	71.67	64.17	65.83	67.50	57.50	61.67	67.50	58.33	69.16	56.67
Water 2	83.85	82.30	85.64	85.64	84.36	83.59	83.08	83.59	84.62	84.62	84.10	82.31
Water 3	82.82	81.29	82.56	81.03	84.10	81.29	83.33	80.76	81.03	80.77	85.39	80.76
Web	58.39	53.02	46.97	55.03	50.37	52.34	42.95	55.70	55.03	57.72	52.34	53.69
Wine	92.70	94.94	95.50	95.50	93.82	91.57	93.82	94.94	94.38	94.38	94.94	93.82

information to be extracted from the fuzzy-rough boundary region for the PART classifier learner. However as this approach only examines the boundary region information, there is no consistency in the results - as can be seen in Fig. 2.

V. CONCLUSIONS

This paper has presented three new techniques for fuzzy-rough feature selection based on the use of fuzzy entropy as an evaluation metric for the fuzzy-rough lower approximations. Note that no user-defined thresholds are required for any of the methods, although a choice must be made regarding fuzzy similarity relations and connectives.

Further work in this area will include a more in-depth experimental investigation of the proposed methods and the impact of the choice of relations and connectives. Additionally, the development of fuzzy discernibility matrices here allows the extension of many existing crisp techniques for the purposes of finding fuzzy-rough reducts. In particular, by reformulating the reduction task in a propositional satisfiability (SAT) framework [2], SAT solution techniques may be applied that should be able to discover such subsets, guaranteeing their minimality. The performance may also be improved through simplifying the fuzzy discernibility function further. This could be achieved by considering the properties of the fuzzy connectives and removing clauses that are redundant in the presence of others.

Also, a more complete comparison of fuzzy-rough feature selection using the metrics proposed in this paper and compared with other FS techniques, would form the basis for a series of topics for future investigation.

REFERENCES

[1] C. Armanino, R. Leardi, S. Lanteri, and, G. Modi Chemom.Intell. Lab.Syst. , vol. 5, pp. 343–354. 1989.

[2] Bayardo, R., and Schrag, R., (1997) Using CSP look-back techniques to solve real-world SAT instances. Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI'97) 203-208

[3] A. Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorisation. Applied Artificial Intelligence, Vol. 15, No. 9, pp. 843–873. 2001.

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

[5] M. Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131-156, 1997.

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

[7] P. Devijver and J. Kittler. Pattern Recognition: A Statistical Approach. Prentice Hall. 1982.

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

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

[10] R. Jensen and Q. Shen, "New Approaches to Fuzzy-Rough Feature Selection," To appear in *IEEE Transactions on Fuzzy Systems*.

[11] N. Mac Parthaláin, R. Jensen, and Q. Shen. Fuzzy Entropy-Assisted Fuzzy-Rough Feature Selection. Proceedings of the 15th International Conference on Fuzzy Systems (FUZZ-IEEE'06). 2006.

[12] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science. 1998.

[13] Z. Pawlak, Rough sets. Int. J. Comput. Inf. Sci. 11, pp. 341–356. 1982.

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

[15] J.R. Quinlan, *C4.5: Programs for Machine Learning*, The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

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

[17] Q. Shen and R. Jensen. Selecting Informative Features with Fuzzy-Rough Sets and its Application for Complex Systems Monitoring. Pattern Recognition, Vol. 37, No. 7, pp. 1351–1363. 2004.

[18] A. Skowron, J. Stepaniuk, Tolerance Approximation Spaces, Fundamenta Informaticae, Vol. 27, pp. 245-253, 1996.

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

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

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