Comparing Fuzzy-Rough and Fuzzy Entropy-assisted Fuzzy-Rough Feature Selection

N. Mac Parthaláin, R. Jensen, and Q. Shen
Dept. of Computer Science.
University of Wales Aberystwyth
Aberystwyth, Ceredigion
WALES
{*nsm03, rkj, qqs*}*@aber.ac.uk*

Abstract

Feature Selection (FS) methods based on fuzzy-rough set theory (FRFS) have employed the dependency function to guide the FS process with much success. More recently a method has been developed which uses fuzzy-entropy [9] to perform this task. Such use of fuzzy-entropy as an evaluation measure in fuzzy-rough feature selection can result in smaller subset sizes than those obtained through FRFS alone. However, it has also been observed that the fuzzy-entropy based FS technique (which does not select subsets based on dependency), also demonstrates remarkably similar dependency values to those of the fuzzy-rough method. This paper investigates the apparent similarity of the dependency values and attempts to discover if any correlation exists. Results are obtained using both fuzzy-rough FS (which is guided solely by the dependency value) and the fuzzy entropy-assisted fuzzy-rough FS technique.

1 Introduction

The task of feature selection is to select a subset of the original features present in a given dataset which provide most of the useful information. Hence, after selection has taken place, most of the important information of the dataset should still remain.In fact, good FS techniques should be able to detect and ignore noisy and misleading features. The result of this, is that dataset quality may even *increase* after selection.

Some of the potential benefits of feature selection include:

1. *Facilitating data visualisation*. By reduction of the data to fewer dimensions, trends within the data can be more easily identified. This can be very important where only a few features have an influence on data outcomes.
2. *Reduction of measurement and storage requirements*. In domains where features correspond to particular measurements (for instance, a water treatment plant [14]), fewer features are highly desirable due to the expense and time-cost of taking such measurements.
3. *Reduction of training and utilisation times*. With smaller datasets, the runtimes of learning algorithms can improve significantly, for both training and classification phases.

This paper examines two FS techniques and compares them on the basis of dependency - a characteristic which is used both as selection and termination criterion in FRFS, whilst fuzzy-rough entropy-assisted feature selection (FREAFS) uses dependency only as a termination criterion - selection being carried out using the entropy value of any considered subset(s).

The principal focus of this paper lies in the comparison of the dependency function values for the above techniques, however an appreciation of both methodologies is necessary in order to gain an understanding for the motivation for this comparison.

The remainder of this paper is structured as follows. Section 2 summarises the theoretical basis and algorithm utilised for FRFS. Section 3 details FREAFS along with a description of the algorithm that is employed for this technique. Section 4 shows the results of applying both FRFS, and FREAFS approaches to a number of datasets (both real and artificially generated), along with a comparison of the dependency values, and level of dimensionality reduction. Section 5 concludes the paper along with suggestions for further work.

2 Fuzzy Rough Feature Selection

A fuzzy-rough set is defined by two fuzzy sets, fuzzy lower and upper approximations, obtained by extending the corresponding crisp rough set notions. In the crisp case, elements that belong to the lower approximation (i.e. have a membership of 1) are said

to belong to the approximated set with absolute certainty. In the fuzzy-rough case, elements may have a membership in the range [0,1], allowing greater flexibility in handling uncertainty.

Fuzzy-Rough Feature Selection [7] is concerned with the reduction of information or decision systems through the use of fuzzy-rough sets. 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 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. For decision systems, $\mathbb{A} = \{\mathbb{C} \cup \mathbb{D}\}$ where $\mathbb{C}$ is the set of input features and $\mathbb{D}$ is the set of decision values.

2.1 Fuzzy Equivalence Classes

Fuzzy equivalence classes [6, 11] are central to the fuzzy-rough set approach in the same way that crisp equivalence classes are central to classical rough sets. 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.

Using the fuzzy similarity relation, the fuzzy equivalence class $[x]_S$ for objects close to x can be defined:

$$\mu_{[x]_S}(y) = \mu_S(x, y) \tag{1}$$

The following axioms should hold for a fuzzy equivalence class F:

- $\exists x$, $\mu_F(x) = 1$
- $\mu_F(x) \wedge \mu_S(x, y) \leq \mu_F(y)$
- $\mu_F(x) \wedge \mu_F(y) \leq \mu_S(x, y)$

The first axiom corresponds to the requirement that an equivalence class is non-empty. The second axiom states that elements in y's neighbourhood are in the equivalence class of y. The final axiom states that any two elements in F are related via the fuzzy similarity relation S. Obviously, this definition degenerates to the normal definition of equivalence classes when S is non-fuzzy. The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes [6].

2.2 Fuzzy Lower and Upper Approximations

The fuzzy lower and upper approximations are fuzzy extensions of their crisp counterparts. Informally, in crisp rough set theory, the lower approximation of a set contains those objects that belong to it with certainty. The upper approximation of a set contains the objects that possibly belong. From the literature, the fuzzy P-lower and P-upper approximations are defined as [6]:

$$\mu_{\underline{P}X}(F_i) = inf_x max\{1 - \mu_{F_i}(x), \mu_X(x)\} \quad \forall i \tag{2}$$

$$\mu_{\overline{P}X}(F_i) = sup_x min\{\mu_{F_i}(x), \mu_X(x)\} \quad \forall i \tag{3}$$

where $\mathbb{U}/P$ stands for the partition of the universe of discourse, $\mathbb{U}$, with respect to a given subset P of features, and F_i denotes a fuzzy equivalence class belonging to $\mathbb{U}/P$. Note that although the universe of discourse in feature reduction is finite, this is not the case in general, hence the use of sup and inf above. 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 redefined as [7]:

$$\mu_{\underline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min(\mu_F(x), \inf_{y \in \mathbb{U}} max\{1 - \mu_F(y), \mu_X(y)\}) \tag{4}$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min(\mu_F(x), \sup_{y \in \mathbb{U}} min\{\mu_F(y), \mu_X(y)\}) \tag{5}$$

The tuple $< \underline{P}X, \overline{P}X >$ is called a fuzzy-rough set. For this particular feature selection method, the upper approximation is not used, though this may be useful for other methods.

For an individual feature, a, the partition of the universe by $\{a\}$ (denoted $\mathbb{U}/IND(\{a\})$) is considered to be the set of those fuzzy equivalence classes for that feature. For example, if the two fuzzy sets N_a and Z_a are generated for feature a during fuzzification, the partition $\mathbb{U}/IND(\{a\}) = \{N_a, Z_a\}$. If the fuzzy-rough feature selection process is to be useful, it must be able to deal with multiple features, finding the dependency between various subsets of the original feature set. For instance, it may be necessary to be able to determine the degree of dependency of the decision feature(s) with respect to feature set $P = \{a, b\}$. In the crisp case, $\mathbb{U}/P$ contains sets of objects grouped together that are indiscernible according to both features a and b. In the fuzzy case, objects may belong to many equivalence classes, so the cartesian product of $\mathbb{U}/IND(\{a\})$ and $\mathbb{U}/IND(\{b\})$ must be considered in determining $\mathbb{U}/P$. In general,

$$\mathbb{U}/P = \otimes\{a \in P : \mathbb{U}/IND(\{a\})\} \quad (6)$$

For example, if $P = \{a, b\}$, $\mathbb{U}/IND(\{a\}) = \{N_a, Z_a\}$ and $\mathbb{U}/IND(\{b\}) = \{N_b, Z_b\}$, then

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

Clearly, each set in $\mathbb{U}/P$ denotes an equivalence class. The extent to which an object belongs to such an equivalence class is therefore calculated by using the conjunction of constituent fuzzy equivalence classes, say F_i, $i = 1, 2, ..., n$:

$$\mu_{F_1 \cap ... \cap F_n}(x) = min(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)) \quad (7)$$

2.3 Fuzzy-Rough Reduction Method

Fuzzy-Rough Feature Selection builds on the notion of the fuzzy lower approximation to enable reduction of datasets containing real-valued features. The process becomes identical to the crisp approach when dealing with nominal well-defined features.

The crisp positive region in the standard RST is defined as the union of the lower approximations. By the extension principle, the membership of an object $x \in \mathbb{U}$, belonging to the fuzzy positive region can be defined by

$$\mu_{POS_P(Q)}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{P}X}(x) \quad (8)$$

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, a new dependency function between a set of features Q and another set P can be defined as follows:

$$\gamma'_P(Q) = \frac{|\mu_{POS_P(Q)}(x)|}{|\mathbb{U}|} = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|} \quad (9)$$

As with crisp rough sets, the dependency of Q on P is the proportion of objects that are discernible out of the entire dataset. In the present approach, this corresponds to determining the fuzzy cardinality of $\mu_{POS_P(Q)}(x)$ divided by the total number of objects in the universe.

A fuzzy-rough QUICKREDUCT algorithm, based on the crisp version [2], has been developed as given in Fig. 1. It employs the fuzzy-rough dependency function γ' to choose which features to add to the current reduct candidate. The algorithm terminates when the addition of any remaining feature does not increase the dependency. As with the original algorithm, for a dimensionality of n, the worst case dataset will result in $(n^2 + n)/2$ evaluations of the dependency function. However, as fuzzy-rough set-based feature selection is used for dimensionality reduction prior to any involvement of the system which will employ those features belonging to the resultant reduct, this operation has no negative impact upon the run-time efficiency of the system.

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

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

Figure 1: The fuzzy-rough QUICKREDUCT algorithm

3 Fuzzy Entropy-Assisted FRFS

Fuzzy Entropy-assisted FRFS uses the FRFS methodology as a basis for dimensionality reduction, while using a fuzzy-entropy measure to guide the FS process, rather than the dependency function value as described in the previous section.

3.1 Classical and Information Entropy (IE)

Classical Entropy may be defined as a measure of the degradation or dispersal of energy and also as the energy form of a system that relates to its internal state of disorder or randomness. Entropy may also be described as a measure of progress of a process of equalisation. It is often used in relation to thermodynamic or metabolic biological processes. High entropy values are indicative of disordered states, and low entropy values are characteristic of ordered states.

Information entropy (IE) or Shannon entropy [13] is also a measure of the amount of disorder in a system and can be defined as:

$$H(X) = -\sum_{i=0}^{N} p_i \log_2 p_i \qquad (10)$$

The entropy of the event *X* is the sum, over all possible outcomes *i* of *X*, of the product of the probability of outcome *i* times the log of the probability of *i*. This can also be applied to a general probability distribution, rather than a discrete-valued event.

The IE value tends to zero with increasing order in any system. It is interesting to note at this point that the fuzzy-rough dependency function value tends to 1 with any increase in order. Having considered this fact, the motivation for investigation of a fuzzy entropy-based approach may not be clear. However, as noted previously, the use of fuzzy-entropy-based techniques often discovered smaller reducts than dependency function-based methods [7].

A fuzzy entropy-assisted approach selects subsets with respect to their entropy value and uses this value to guide the feature selection process.

3.2 Fuzzy Entropy Measure

Again, let $I = (\mathbb{U}, \mathbb{A})$ be a decision system, where $\mathbb{U}$ is a non-empty set of finite objects. $\mathbb{A} = \{\mathbb{C} \cup \mathbb{D}\}$ is a non-empty finite set of attributes, where $\mathbb{C}$ is the set of input features and $\mathbb{D}$ is the set of classes. An attribute $a \in \mathbb{A}$ has corresponding fuzzy subsets $F_1, F_2, ..., F_n$. The fuzzy entropy for a fuzzy subset F_i can be defined as:

$$H(F_i) = -\sum_{D \in \mathbb{U}/\mathbb{D}} p(D|F_i) \log_2 p(D|F_i) \qquad (11)$$

where, $p(D|F_i)$ is the relative frequency of the fuzzy subset F_i of attribute a with respect to the decision D, and is defined:

$$p(D|F_i) = \frac{|D \cap F_i|}{|F_i|} \qquad (12)$$

The cardinality of a fuzzy set is denoted by $|\cdot|$. Based on these definitions, the fuzzy entropy for an attribute subset R is defined as follows:

$$E(R) = \sum_{F_i \in \mathbb{U}/R} \frac{|F_i|}{\sum_{Y_i \in \mathbb{U}/R} |Y_i|} H(F_i) \qquad (13)$$

This fuzzy entropy can be used to gauge the utility of attribute subsets in a similar way to that of the fuzzy-rough measure. However, the fuzzy entropy measure decreases with increasing subset utility, whereas the fuzzy-rough dependency measure increases. With these definitions, a new feature selection mechanism can be constructed that uses fuzzy entropy to guide the search for the best fuzzy-rough feature subset.

3.3 Fuzzy-Rough Entropy-based QUICKREDUCT

Figure 2 below shows a fuzzy-rough entropy-based QUICKREDUCT algorithm based on the previously described fuzzy-rough algorithm in figure 1.

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

(1) $T \leftarrow \{\}, \gamma'_{prev} \leftarrow 0$
(2) **do**
(3) $R \leftarrow T$
(4) $\gamma'_{prev} \leftarrow \gamma'_T(\mathbb{D})$
(5) $\forall x \in (\mathbb{C} - R)$
(6) **if** $E(R \cup \{x\}) < E(T)$
(7) $T \leftarrow R \cup \{x\}$
(8) **until** $\gamma'_T(\mathbb{D}) \leq \gamma'_{prev}$
(9) **return** R

Figure 2: The fuzzy-rough fuzzy entropy-based QUICKREDUCT algorithm

FREQUICKREDUCT is similar to the fuzzy-rough algorithm but uses the entropy value of a data subset to guide the feature selection process. If the fuzzy entropy value of the current reduct candidate is smaller than the previous, then this reduct is retained and used in the next iteration of the loop. It is important to point out that the reduct is evaluated by examining its entropy value, termination only occurs when the addition of any remaining features n results in a decrease in the dependency function value (γ'_{prev}). The fuzzy-entropy value therefore is not used as a termination criteria.

The algorithm begins with an empty subset R and with γ'_{prev} initialised to zero. The do-until loop works by examining the entropy value of a subset and incrementally adding one conditional feature at a time, until the dependency function value begins to fall to a value that is lower or equal to that of the last subset. For each iteration, a conditional feature that has not already been evaluated will be temporarily added to the subset R. The entropy of the subset currently being examined (5) is then evaluated and compared with the entropy of T, (the previous subset). If the entropy value of the current subset is lower (6), then the attribute added in (5) is retained as part of the new reduct T (7).

The loop continues to evaluate in the above man-

ner by adding conditional features, until the dependency value of the current reduct candidate ($\gamma'_R(\mathbb{D})$) falls to a value lower than or equal to that of the previously evaluated reduct candidate.

3.4 A Worked Example

To illustrate the operation of the new fuzzy entropy-based algorithm, a small example dataset (given in table **??**) is considered, containing real-valued conditional attributes with nominal decisions.

Table **??** contains three real-valued conditional attributes and a crisp-valued decision attribute. To begin with, the algorithm initializes the potential reduct (i.e. the current best set of attributes) to the empty set.

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

Table 1: Example dataset: crisp decisions

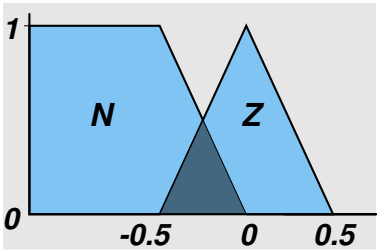

Figure 3: Fuzzifications for conditional features

Using the fuzzy sets defined in figure **??** (for all conditional attributes), and setting $A = \{a\}$, $B = \{b\}$, $C = \{c\}$ and $\mathbb{D} = \{q\}$, the following equivalence classes are obtained:

$$\begin{aligned}\mathbb{U}/A &= \{N_a, Z_a\}\\ \mathbb{U}/B &= \{N_b, Z_b\}\\ \mathbb{U}/C &= \{N_c, Z_c\}\\ \mathbb{U}/\mathbb{D} &= \{\{1,3,6\},\{2,4,5\}\} = \{D_1, D_2\}\end{aligned}$$

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 ($A = \{a\}$):

$$E(A) = \frac{|N_a|}{|N_a| + |Z_a|} H(N_a) + \frac{|Z_a|}{|N_a| + |Z_a|} H(Z_a)$$

For the first part of the summation, the value $H(N_a)$ must be determined. This is achieved in the following way:

$$\begin{aligned}H(N_a) &= -\textstyle\sum_{D \in \mathbb{U}/\mathbb{D}} p(D|N_a) \log_2 p(D|N_a)\\ &= -p(D_1|N_a) \log_2 p(D_1|N_a)\\ &\quad + -p(D_2|N_a) \log_2 p(D_2|N_a)\end{aligned}$$

The required probabilities are $p(D_1|N_a) = 0.6363637$, $p(D_2|N_a) = 0.3636363$. Hence, $H(N_a) = 0.94566023$. In a similar way, $H(Z_a)$ can be calculated, giving a value of 1.0.

To determine the fuzzy entropy for a, the values $\frac{|N_a|}{|N_a|+|Z_a|}$ and $\frac{|Z_a|}{|N_a|+|Z_a|}$ must also be determined. This is achieved through the standard fuzzy cardinality, resulting in a fuzzy entropy value of:

$$\begin{aligned}E(A) &= (0.47826084.H(N_a)) + (0.5217391.H(Z_a))\\ &= (0.47826084 \times 0.94566023)\\ &\quad + (0.5217391 \times 1.0)\\ &= 0.9740114\end{aligned}$$

Repeating this process for the remaining attributes gives:

$$\begin{aligned}E(B) &= 0.99629750\\ E(C) &= 0.99999994\end{aligned}$$

From this it can be seen that attribute a will cause the greatest decrease in fuzzy entropy. This attribute is chosen and added to the potential reduct, $R \leftarrow R \cup \{a\}$. This subset is then evaluated using the fuzzy-rough dependency measure, resulting in $\gamma_R(\mathbb{D}) = 0.3333333$. The previous dependency value is 0 (the algorithm started with the empty set), hence the search continues. The process iterates and the two fuzzy entropy values calculated are

$$\begin{aligned}E(\{a,b\}) &= 0.7878490\\ E(\{a,c\}) &= 0.9506136\end{aligned}$$

Adding attribute b to the reduct candidate causes the larger decrease of fuzzy entropy, so the new candidate becomes $\{a,b\}$. The resulting dependency value for this, $\gamma_{\{a,b\}}(\mathbb{D})$, is 0.56666666. This is, again, larger than the previous dependency value, and so search continues. Lastly, attribute c is added to the potential reduct:

$$\begin{aligned}E(\{a,b,c\}) &= 0.7412282\\ (\gamma_{\{a,b,c\}}(\mathbb{D}) &= 0.56666666)\end{aligned}$$

As this causes no increase in dependency, the algorithm stops and outputs the reduct $\{a,b\}$. The dataset can now be reduced to only those attributes appearing in the reduct.

4 Experimentation

This section presents the results of the experimental studies using the datasets described in table 1. These datasets are small-to-medium in size, with between 8 and 390 objects. Also included are some datasets which have been generated artificially which have between 2 and 5 decision classes. These artificial datasets have randomly generated condition attributes (that are a mix of both real and crisp values), whilst the decision attributes are based on an inequality function for each dataset which relates to the values of conditional attributes. This allows the decision attributes to be manipulated and hence is a good indicator of the efficiency of the employed FS technique. A good FS technique should select only the attributes that are part of the inequality expression. More information on the real data used in this paper can be found in [10] and also [9]

A comparison of the entropy and FRFS-based dimensionality reduction techniques is given based on the dependency values obtained for each approach.

4.1 Dependency Function

Since the principal focus of this paper lies in examining the dependency function value of both the fuzzy-rough and entropy-assisted fuzzy-rough approaches to FS, it is important to note that the entropy-assisted approach as described in [9] selects reducts for consideration on the basis of entropy value for that reduct.

Table 2 presents a comparison of reduct size, and dependency value, using both FRFS and entropy-based approaches.

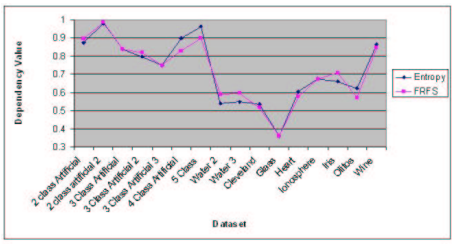

Figure 4: Dependency Function value: FRFS and FEAFRFS

It is clear from the results obtained that both FRFS and entropy-assisted methods both reflect very similar (in some cases identical) values of dependency as noted in [9]. Indeed when comparing all of these values, it becomes clear that there is only a maximum difference in the order of 0.06.

As mentioned previously, the entropy-assisted approach can result in reducts which are smaller than those obtained using FRFS. However the results show there is no direct relationship between reduct size and dependency value. In fact, there are instances where the dependency value has fallen or remained at very similar values for the FEAFRFS approach that have still resulted in subsets that are smaller than those obtained by FRFS.

Considering the results for the artificial data, it is apparent that no significant decrease in reduct size is obtained for FEAFRFS, although this is to be expected as the data has been manually manipulated. However, on closer examination of the results obtained for this data it was discovered although the reducts were of the same size, the attributes selected were not always the same. Also when compared with the inequality used for the decision attribute it was discovered that the FEAFRFS reduct always succeeded in including the correct attributes. This is an interesting aspect as it suggests that the FEAFRFS approach is a better mechanism for FS.

4.2 Dimensionality Reduction

From Fig.4 it is apparent that there is no correlation between the level of dimensionality reduction and the dependency level for FEAFRFS. Indeed there are some cases where the dependency of the FEAFRFS approach has a lower value than that of FRFS yet still manages to produce a better reduction in dimensionality. Also the artificial data shows that although both approaches achieved the same level of dimensionality reduction, FEAFRFS was able to return higher levels of dependency.

4.3 Correlation of Dependency Values

When considering the dependency data for both approaches in table 2, it was noted that the values were remarkably similar. Further investigation revealed that using the Pearson correlation coefficient (PMCC), a value of 0.9786794 was obtained (a value of 1 is perfect correlation) thus indicating that both values are highly positively correlated. Also, it is interesting to note that in those cases where FEAFRFS succeeds in finding smaller reducts than FRFS, correlation tends to be even higher(0.9847) between dependency values.

5 Conclusions

The dependency function values produced by FEAFRFS method very closely follow those of the FRFS method. It has been observed that there is no relation between the dependency value and the level of dimensionality reduction obtained.

The correlation of the dependency values for

Dataset Name	Objects	Features	Decision feat. type	Description
art1	8	5	binary	artificially generated dataset
art2	8	12	binary	artificially generated dataset
art3	20	9	3-class	artificially generated dataset
art4	6	5	3-class	artificially generated dataset
art5	150	5	3-class	artificially generated dataset
art6	8	16	4-class	artificially generated dataset
art7	14	20	5-class	artificially generated dataset

Table 2: Artificial Data

Dataset	Original number of features	Reduct size		Final dependency value	
		FRFS	Entropy	FRFS	Entropy
water 2	39	11	8	0.588	0.540
water 3	39	12	11	0.595	0.549
cleveland	14	11	10	0.516	0.535
glass	10	9	9	0.359	0.359
heart	14	11	9	0.578	0.607
ionosphere	35	11	11	0.673	0.677
iris	5	5	3	0.707	0.658
olitos	26	10	8	0.572	0.620
wine	14	10	9	0.844	0.862
art1	5	3	2	0.895	0.871
art2	12	3	3	0.983	0.978
art3	9	5	5	0.838	0.838
art4	5	4	4	0.820	0.798
art5	5	5	5	0.748	0.748
art6	15	4	4	0.828	0.898
art7	20	8	7	0.899	0.961

Table 3: Comparison of Dependency values & Reduct size

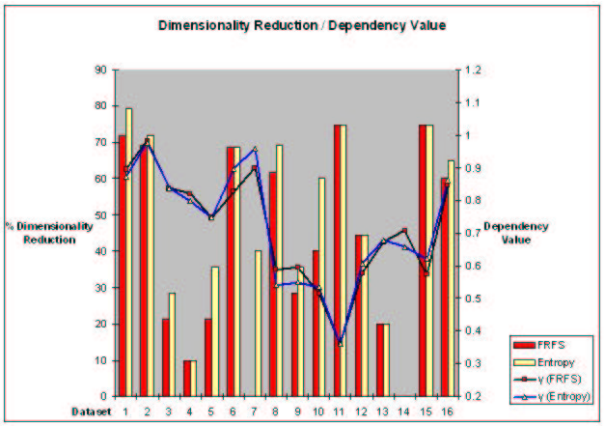

Figure 5: Dimensionality Reduction vs. Dependency Function value

each approach is very high generally (0.9786794) and this correlation is even higher in instances where FEAFRFS manages better reducts than the corresponding FRFS results.

An important observation was made when analysing the artificial data, this related to the fact that FEAFRFS always selected the correct features in relation to the inequality. This highlights the superiority of FEAFRFS as a feature selector despite the fact that this approach is computationally more complex.

An area of interest that has not been explored in this paper, but one that may reveal some interesting results is the comparison of the dependency versus the fuzzy entropy for each attribute that is considered for selection. This would highlight any correlation that may exist between these metrics and may provide scope for further work.

References

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

[2] A. Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorisation. Ap-

plied Artificial Intelligence, Vol. 15, No. 9, pp. 843–873. 2001.

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

[4] S. M. Chen and J. D. Shie, A new method for feature subset selection for handling classification problems, Proceedings of the 2005 IEEE International Conference on Fuzzy Systems, Reno, Nevada, US, pp. 183–188, May 2005.

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

[6] D. Dubois and H. Prade. Putting rough sets and fuzzy sets together. In [15], pp. 203–232. 1992.

[7] R. Jensen and Q. Shen. Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches. IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 12, pp. 1457–1471. 2004.

[8] B. Kosko. Fuzzy entropy and conditioning. Information Sciences, Vol. 40, No. 2, pp. 165–174. 1986.

[9] 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.

[10] 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.

[11] S.K. Pal and A. Skowron (Eds.). Rough-Fuzzy Hybridization: A New Trend in Decision Making. Springer Verlag, Singapore. 1999.

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

[13] C.E. Shannon, A mathematical theory of communication. Bell System Technical Journal, vol. 27, pp. 379–423, and pp. 623-656, July and October, 1948.

[14] 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.

[15] R. Slowinski, editor. Intelligent Decision Support. Kluwer Academic Publishers, Dordrecht. 1992.

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

[17] 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.