2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

Feature Grouping-Based Fuzzy-Rough Feature Selection

Richard Jensen
Dept. of Computer Science
Aberystwyth University
Aberystwyth, Ceredigion, Wales, UK
Email: rkj@aber.ac.uk

Neil Mac Parthaláin
Dept. of Computer Science
Aberystwyth University
Aberystwyth, Ceredigion, Wales, UK
Email: ncm@aber.ac.uk

Chris Cornelis
Dept. of Computer Science and AI
CITIC-UGR, University of Granada
Granada, Spain
Email: chris.cornelis@decsai.ugr.es

***Abstract*—Data dimensionality has become a pervasive problem in many areas that require the learning of interpretable models. This has become particularly pronounced in recent years with the seemingly relentless growth in the size of datasets. Indeed, as the number of dimensions increases, the number of data instances required in order to generate accurate models increases exponentially. Feature selection has therefore become not only a useful step in the process of model learning, but rather an increasingly necessary one. Rough set and fuzzy-rough set theory have been used as such dataset pre-processors with much success, however the underlying time/space complexity of the subset evaluation metric is an obstacle to the processing of very large data. This paper proposes a general approach to this problem that employs a novel feature grouping step in order to alleviate the processing overhead for large datasets. The approach is framed within the context of (and applied to) fuzzy-rough sets, although it can be used with other subset evaluation techniques. The experimental evaluation demonstrates that considerable computational effort can be avoided, and as a result efficiency can be improved considerably for larger datasets.**

***Index Terms*—fuzzy-rough sets, feature selection, feature grouping.**

I. Introduction

The impact of data abundance now extends well beyond the traditional areas of concern such as machine learning. Indeed, in recent years, in fields and disciplines as varied as science, sports, public health and even advertising, there is a move toward data-driven knowledge discovery and decision-making. However, this unrelenting drive towards quantification and wealth of new data only encourages the further archiving of enormous amounts of data. This seemingly vicious cycle is the impetus for the development of many of the techniques which aim to reduce the size of data to a form which is more compact, more interpretable, and more computationally tractable. Many real-world problems involve the specification of high dimensional descriptions of input feature spaces. It is not surprising therefore that much research has been carried out in the area of dimensionality reduction and feature selection [3]. However, much of that existing work tends to destroy the underlying semantics of the features after reduction or requires additional meta-information about the supplied data for thresholding. A technique that can reduce dimensionality using the information contained within the dataset only and that simultaneously preserves the meaning of the features (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 [14], [15].

Over the past 15 years, RST has attracted much interest from researchers and has been applied to many domains. 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 very little information loss. Therefore, there has been much research in the area of finding reducts, and this has since been extended to fuzzy-rough feature selection in order to handle data with real-valued features [12], [9]. Such approaches, whilst powerful, rely on a subset evaluation metric which is relatively expensive from a computational standpoint, and can become prohibitive for larger datasets particularly when the data contains a large number of features which are highly correlated with one another.

When considering data of large dimensionality, much computational time may be expended in examining features that are strongly correlated with each other (i.e. have high levels of redundancy), and hence carry similar information. Currently, no mechanism exists in order to consider this type of situation in the fuzzy-rough framework, which results in much wasted computational effort. In order to combat this, an approach is described here that groups correlated features, prior to the feature selection phase as a pre-processing procedure. The process of feature selection is then carried out on the basis of the groups which have previously been formed. Such a procedure not only reduces the amount time taken to process large data but also has the potential to generate feature subsets of better quality with lower levels of internal redundancy.

The remainder of this paper is structured as follows. Firstly, the key concepts that underpin rough and fuzzy-rough set theory are reviewed. Next, the new approach for feature grouping in feature selection is presented with a simple worked example in order to illustrate the overall process. The results of an experimental evaluation based on some benchmark problems are then presented, where the approach is compared to some current state of the art techniques. Finally the paper

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

is concluded and some conclusions are drawn.

II. Theoretical Background

A. Rough set theory

At the very heart of the RST is the concept of indiscernibility [14]. Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite instances (the universe of discourse) and $\mathbb{A}$ is a non-empty finite set of features 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 decision systems (the focus of the rest of this paper), $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$ where $\mathbb{C}$ is the set of input features and $\mathbb{D}$ is the set of decision features.

For any $P \subseteq \mathbb{C}$, 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 $\mathbb{U}/P$ and is calculated as follows:

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

where,

$$S \otimes T = \{X \cap Y : \forall X \in S, \forall Y \in T, X \cap Y \neq \emptyset\} \quad (3)$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]_P$. 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:

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

$$\overline{P}X = \{x : [x]_P \cap X \neq \emptyset\} \quad (5)$$

The positive region can then be constructed, which contains those objects for which the values of P allow to predict the decision classes unequivocally:

$$POS_P(\mathbb{D}) = \bigcup_{X \in \mathbb{U}/\mathbb{D}} \underline{P}X \quad (6)$$

By employing this definition of the positive region it is possible to calculate the rough set degree of dependency of the decision attribute(s) $\mathbb{D}$ on a set of attributes P:

$$\gamma_P(\mathbb{D}) = \frac{|POS_P(\mathbb{D})|}{|\mathbb{U}|} \quad (7)$$

B. Fuzzy-rough set theory

Fuzzy-rough sets have been proposed in order to improve the ability to deal with uncertainty and vagueness present in data. A fuzzy-rough set [5] is defined by two fuzzy sets, fuzzy lower and upper approximations, obtained by extending the corresponding crisp rough set notions defined in (4) and (5) previously. 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], thus allowing greater flexibility in handling uncertainty. Fuzzy-rough sets encapsulate the related but distinct concepts of vagueness (for fuzzy sets) and indiscernibility (for rough sets), both of which occur as a result of uncertainty in knowledge.

Definitions for the fuzzy lower and upper approximations can be found in [17], where a fuzzy indiscernibility relation is used to approximate a fuzzy concept X:

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in \mathbb{U}} \mathcal{I}(\mu_{R_P}(x, y), \mu_X(y)) \quad (8)$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in \mathbb{U}} \mathcal{T}(\mu_{R_P}(x, y), \mu_X(y)) \quad (9)$$

Here, $\mathcal{I}$ is a fuzzy implicator and $\mathcal{T}$ a t-norm. A fuzzy implicator is any $[0, 1]^2 \rightarrow [0, 1]$-mapping $\mathcal{I}$ satisfying $\mathcal{I}(0, 0) = 1, \mathcal{I}(1, x) = x$ for all x in $[0, 1]$. R_P is the fuzzy similarity relation induced by the subset of features P:

$$\mu_{R_P}(x, y) = \mathcal{T}_{a \in P}\{\mu_{R_a}(x, y)\} \quad (10)$$

$\mu_{R_a}(x, y)$ is the degree to which instances x and y are similar for feature a, and may be defined in many ways, for example:

$$\mu_{R_a}(x, y) = 1 - \frac{|a(x) - a(y)|}{|a_{max} - a_{min}|} \quad (11)$$

$$\mu_{R_a}(x, y) = \exp(-\frac{(a(x) - a(y))^2}{2{\sigma_a}^2}) \quad (12)$$

$$\mu_{R_a}(x, y) = \max(\min(\frac{(a(y) - (a(x) - \sigma_a))}{\sigma_a}, \frac{((a(x) + \sigma_a) - a(y))}{\sigma_a}, 0) \quad (13)$$

where ${\sigma_a}^2$ is the variance of feature a. The choice of relation is largely determined by the intended application. For feature selection, a relation such as (13) may be appropriate as this permits only small differences between attribute values of differing instances. For classification tasks, a more gradual and inclusive relation such as (11) should be used.

In a similar way to the original crisp rough set approach, the fuzzy positive region [12] can be defined as:

$$\mu_{POS_P(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{R_P}X}(x) \quad (14)$$

An important issue in data analysis is discovering dependencies between attributes. The fuzzy-rough degree of dependency of $\mathbb{D}$ on the attribute subset P can be defined in the following way:

$$\gamma'_P(\mathbb{D}) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_P(\mathbb{D})}(x)}{|\mathbb{U}|} \quad (15)$$

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 greedy hill-climbing algorithm can be constructed that uses equation (15) to gauge subset quality. In [12], it has been shown that the dependency function is monotonic and that

fuzzy discernibility matrices may also be used to discover reducts.

III. Feature grouping-based selection

One of the primary impediments associated with conventional greedy hill-climbing approaches to discovering fuzzy-rough reducts in large datasets is that much time is wasted considering features that have strong correlation with each other. The consideration of such features is somewhat superfluous as they contain very similar information. Ultimately, evaluating all such features at each stage of the search offers no advantage. Take for example, an extreme situation where a particular dataset contains several hundred replicated features. A hill-climbing type of search will consider the addition of each of these features to the current subset candidate iteratively at each stage of the search. Obviously, such computation is completely unnecessary. Furthermore, the later addition of any features to the subset candidate will often produce only very small improvements in the overall fuzzy-rough dependency metric [2]. The result of which is a super-reduct, i.e. the resulting subset contains superfluous features that are redundant and can be otherwise removed with no loss in dependency.

The approach proposed here (abbreviated FRFG hereafter), aims to group together similar features such that at each stage of hill-climbing, only the most promising group representative is considered for selection. This will reduce wasted computational effort, and also help to improve the final selected subset quality.

A. Forming groups

An important component of the proposed approach is the identification of related features and how groups of same are formed appropriately. There are many measures that are useful for this task. Here, the sample correlation coefficient is used:

$$corr(a,b) = \frac{\sum_{i=1}^{|\mathbb{U}|}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{|\mathbb{U}|}(a_i - \overline{a})^2 \sum_{i=1}^{|\mathbb{U}|}(b_i - \overline{b})^2}} \quad (16)$$

where $a, b \in \mathbb{A}$, and $\overline{a}$ refers to the sample mean of a. This measure can be used to evaluate the degree of correlation between conditional features in order to determine groups. The sample correlation coefficient ranges from -1 to +1. In this work, the absolute value is used as a feature that is negatively correlated with another feature can also be considered to be redundant:

$$correlation(a,b) = |corr(a,b)| \quad (17)$$

The same correlation measure can be used to evaluate the correlation of conditional features with the class attribute in order to rank features within these groups. The most relevant features (according to the correlation measure) are ranked highest in the groups. It is from these groups that the adapted hill-climbing method will select features. Redundancy is therefore partly handled by employing groups of similar features, and relevance is considered by ranking features within groups based on their relatedness to the decision feature.

Having calculated the correlations values, groups can then formed. Here, a threshold is used to determine group membership of features. This threshold could be either a user-supplied (τ), that must be exceeded for a pair of features to be considered redundant, or could be estimated automatically:

$$\tau = 0.8(\max_{a,b \in \mathbb{C}}\{correlation(a,b)\}) \quad (18)$$

Groups are formed in the following way. For each feature f_i, the correlation with every other feature f_j is determined and the threshold (τ) applied such that if the correlation is greater than the threshold, then the feature f_j is added to the group for f_i, i.e. $F_i \leftarrow F_i \cup \{f_j\}$. Having considered all features, the result is a set of groups $F = \{F_1, F_2, ..., F_{|\mathbb{C}|}\}$. Features can be ordered within groups on the basis of their correlation with the decision feature $\mathbb{D}$, meaning that features that have greater correlation with $\mathbb{D}$ are preferable. It is important to note that as a result of this process, features can belong to more than one group.

B. Subset search

Having formed the groups, the next phase of the FRFG approach is to employ the groups and their respective internal rankings in order to guide the search procedure in discovering good subsets according to a given metric. In this paper, the fuzzy-rough dependency measure is used to gauge subset quality, however any measure can be used for this purpose (including wrapper approaches). As mentioned previously, an adapted hill-climbing algorithm is used here to find the best subsets. Although there are some issues with greedy approaches (e.g. see [13]), it is still a useful search mechanism and often discovers reducts or superreducts that are usually only slightly larger than optimal. The way in which the hill-climbing search is formulated means that it is reasonably straightforward to reconfigure it for a group-based strategy. The full algorithm can be seen in Figure 1, including the required initialization steps.

The purpose of the function preprocess(F) is to perform some initial pre-processing in order to investigate if there is any perfect correlation between features, and to remove the less relevant feature each time. This could be softened to use another threshold to remove more features (i.e. for threshold values less than 1), however this may remove useful features and prevent the algorithm from finding an optimal reduct.

For each group of features, the representative top-ranked feature is chosen and assessed by temporarily adding it to the current reduct candidate and evaluating this new subset via the metric M. In this paper, the focus is fuzzy-rough feature selection, and hence the measure used for M is the fuzzy-rough dependency degree. Once a feature has been evaluated, its group members are then added to the *Avoids* set to ensure that these features are not evaluated in this iteration. The feature that produces the greatest increase in the metric is then added to the current subset and the process iterates until

the stopping criterion is fulfilled. This may involve stopping when the maximum value for the measure has been reached, or to degree α, or indeed if there is no change in the measure following two successive iterations. In the fuzzy-rough case, the maximum value for a dataset can be determined prior to selection and then used as a stopping criterion.

Line (14) provides an optional further reduction in computational effort (set by the Boolean flag '`moreAvoids`') by removing all other features which appear in the group of that newly selected feature from consideration. The rationale for this step is that once a feature has been selected, the addition of any of its group members at this stage will not benefit the overall subset. There may be *some* utility in allowing the possibility of correlated group members to be added [7], but it is unlikely to have great impact on the evaluation metric. However, for flexibility, the addition of other group members of previously selected features can be permitted if this flag is set to `false`. In this work, the default setting is `true`.

In the extreme case, by setting the threshold τ=0, the algorithm then acts as a ranking approach that adds features to the reduct candidate linearly on the basis of their relevance, until the subset evaluation measure has been maximized. However, if `moreAvoids` is set to true, this behaviour will not be exhibited; instead, only the first, most relevant, feature will be chosen and then the algorithm will terminate (all other features appear in its group and are therefore removed from consideration).

FRFG($\mathbb{C}$, $\mathbb{D}$, M,τ,`moreAvoids`).
$\mathbb{C}$, the set of conditional features;
$\mathbb{D}$, the set of decision features;
M, subset evaluation measure;
τ, the group-forming threshold;
`moreAvoids`, Boolean variable

(1) $R \leftarrow \emptyset$; $F \leftarrow$ formGroups($\mathbb{C}, \tau$)
(2) $F \leftarrow$ rankWithinGroups($\mathbb{D}$,F)
(3) preprocess(F); $F \leftarrow$ order(F); $AlwaysAv \leftarrow \emptyset$
(4) **while** (stopping criterion not met)
(5) $\quad Avoids \leftarrow AlwaysAv$
(6) $\quad$ bestF $\leftarrow \emptyset$; bestEval $= 0$
(7) $\quad$ **foreach** $a \in (\mathbb{C} - R - Avoids)$
(8) $\quad\quad a \leftarrow$ highestRankedFeature(F_a)
(9) $\quad\quad T \leftarrow R \cup \{a\}$
(10) $\quad\quad$ **if** $(M(T) >$bestEval)
(11) $\quad\quad\quad$ bestF = a; bestEval = $M(T)$
(12) $\quad\quad Avoids \leftarrow Avoids \cup F_a$
(13) $\quad R \leftarrow R \cup$ bestF
(14) $\quad$ **if** (`moreAvoids`)
(15) $\quad\quad AlwaysAv \leftarrow AlwaysAv \cup F_{bestF}$
(16) **output** R

Fig. 1: The feature grouping algorithm

The function order(F) orders the considered feature groups on the basis of their top-ranked features (i.e. most relevant), so the most promising groups are considered first. Without this, the algorithm may favour earlier features in an arbitrary fashion.

Once a feature has been added to the current subset, its group members are removed from consideration at this level. However, this does not prevent consideration of this group in future iterations. The search will stop when the stopping criterion is met. For many filter measures, a known maximum is attainable and therefore this is used to judge when to terminate the algorithm. For other measures, search can be halted when there is little or no perceived improvement in the subset quality. Also, it may be useful to stop the search somewhat prematurely by using a threshold, α, that indicates when a subset is *good enough*.

C. Worked Example

To illustrate the FRFG approach, an artificial example is described here. Consider a dataset with six features, some of which are highly correlated. After the initialization steps of the algorithm, the groups formed are:

$$\begin{aligned} F_1 &= \{f_4, f_3, f_1\} \\ F_2 &= \{f_2\} \\ F_3 &= \{f_3, f_1\} \\ F_4 &= \{f_4, f_1, f_5\} \\ F_5 &= \{f_4, f_5\} \\ F_6 &= \{f_6\} \end{aligned}$$

Here, features within the groups have been ordered according to their relevance, so the left-most features are more relevant to the decision and thus are preferable to those on the right. Groups F_2 and F_6 have only one member, which indicates that features f_2 and f_6 are not strongly correlated with other features.

The hill-climbing algorithm first orders the group, say $F = \{F_4, F_3, F_1, F_5, F_2, F_6\}$ and begins the search at the first level. The first group to be considered is F_4; feature f_4 is preferable over others and is therefore added to the current (initially empty) subset R . This is then evaluated: $M(R \cup \{f_4\})$ and if it results in a better score than the current best evaluation, then feature f_4 is stored and the current best evaluation is set to $M(R \cup \{f_4\})$. The set of features which appears in group F_4 is then added to the set *Avoids* so that other group members are not evaluated in this iteration. In other words, once the main group representative has been selected, other highly correlated group members do not need to be considered. Therefore, $Avoids = \{f_1, f_4, f_5\}$ and the next feature group is considered that does not appear in *Avoids*, F_3. The highest ranked feature, f_3, is then added to the current subset and evaluated, $M(R \cup \{f_3\})$. If this value is greater than $M(R \cup \{f_4\})$, then feature f_3 replaces f_4. The set *Avoids* is then updated with the members of F_3, $Avoids = \{f_1, f_3, f_4, f_5\}$.

The next feature groups F_1 and F_5 both appear in *Avoids* and so are not considered. This means that the next considered

group is F_2 (which consists of a single feature) is evaluated. Finally, the single remaining group F_6 is considered and evaluated. Having completed this, the best representative feature in this iteration is then added to the reduct candidate R and the process iterates once more (unless the stopping criterion is not met). The *Avoids* list is reset at this level.

From this small example, it can be seen that considerable computational effort has been avoided since features f_1 and f_5 did not need to be evaluated. Note that the level of computational effort saved is governed by the group sizes, which in turn is decided by the thresholding which is used in order to form the groups. Hence, a balance must be maintained between lower thresholds (which produce larger groups, greater time saving, but potentially group fewer correlated features together) and higher thresholds (which produce smaller groups, less time saving, but features within groups are more highly correlated). In the extreme case where the threshold is set to 1, the algorithm becomes a standard hill-climber where each feature appears in exactly one group, and no time saving is made during execution. The worst-case complexity of this is O($|\mathbb{C}|^2$). In the other extreme, where the threshold is set to 0, all features are grouped together in ranked order and the selection process simply selects features based upon their ranking (derived form the correlation metric) until the stopping criterion is met. The worst-case complexity in this situation is linear in the number of features, O($|\mathbb{C}|$). Depending on the threshold value employed therefore, the actual worst-case complexity will lie somewhere between quadratic and linear for a given dataset.

IV. Experimental Evaluation

This section details the experiments conducted and the results obtained for the novel FRFG approach. In a series of experiments, the proposed approach was applied to 10 datasets of different sizes, and compared with three other search methods for discovering fuzzy-rough reducts. The results presented here relate to performance in terms of quality of subsets obtained: classification accuracy and subset size, as well as execution times, and the effect of a range of threshold values for τ on the results for the FRFG approach.

1) Experimental Setup: The experimentation employed a total of 10 different datasets, which are detailed in table I. Eight of these datasets are drawn from [6], whilst the remaining two are real-world mammographic risk-assessment tasks which are related to data derived from [19] and [8] and features extracted/decision class labelling schemes from the work in [10].

For the purposes of comparison, three approaches for feature selection are also included. All of of these use the fuzzy-rough subset evaluation metric as described in [12], along with three different reduct search methods: greedy hill-climbing (GHC), GA search, and PSO search. In addition, five different experiments are carried out for the novel FRFG approach by imposing different values for the threshold τ: 0.0, 0.2, 0.4, 0.6, 0.8 and 0.9. Note that for the experimentation with τ=0, `moreAvoids` is set to `false`; in this case, the algorithm will add features in order of rank to the reduct candidate until the fuzzy-rough dependency has reached its maximal value.

Dataset	Features	Instances
MIAS	281	322
DDSM	281	832
web	2557	149
cleveland	13	297
glass	9	214
heart	13	270
olitos	25	120
water2	39	390
water3	39	390
wine	13	178

TABLE I: Benchmark data

The GA-based search has an initial population size of 200, the maximum number of generations was set to 40, with the crossover probability set to 0.6 and mutation probability set to 0.033. The number of generations in the case of the PSO-based search was also set to 40, whilst the number of particles was set to 200, with acceleration constants $c1$ = 1 and $c2$ = 2.

For the fuzzy-rough subset evaluation metric, the Łukasiewicz t-norm ($\max(x+y-1,0)$) and the Łukasiewicz fuzzy implicator ($\min(1-x+y,1)$) are adopted to implement the fuzzy connectives in (8) and (9).

For the generation of classification results, three different classifier learners have been employed: J48 which is based on ID3 [16]; JRip, a rule-based classifier [1]; and IBk, a nearest-neighbour classifier (with k = 3). Five stratified randomisations of 10-fold crossvalidation were employed in generating the classification results. It is important to point out here that feature selection is performed as part of the crossvalidation and each fold results in a new selection of features. Finally for the comparison of FRFG with the other approaches in terms of classification accuracy, a statistical significance test was performed using a corrected paired t-test (significance value: 0.05) in order to ensure that the results obtained were statistically significant.

2) Results: The results of the experimental evaluation are shown in tables II - VI. Tables II – IV detail the classification results for the J48, JRip and IBk classifier learners respectively. GHC (greedy hill-climbing), GA (genetic algorithm) and PSO (particle swarm optimisation) refer to the search technique employed in each case. Examining these results, it is clear that regardless of the value of τ, FRFG returns very similar results to GHC. Indeed, when a paired t-test is employed to examine the statistical significance of the results generated for FRFG, only those results for the *wine* dataset where τ=0.2 and 0.4 are statistically inferior to those for GHC. It is worth noting from table V however, that the average subset size for these values of τ, is much smaller than for GHC indicating a trade-off between compactness of representation and model accuracy.

When FRFG is compared with the GA-based search, a similar pattern emerges. However, in this case, FRFG does not return any results which are statistically inferior. It is the

Dataset	Unred.	GHC	FRFG τ =						GA	PSO
			0.0	0.2	0.4	0.6	0.8	0.9		
MIAS	66.72	60.11	57.22	61.98	62.99	61.42	61.63	59.26	61.88	52.67
DDSM	50.16	46.40	44.24	50.69	46.86	47.20	45.48	46.43	48.71	47.39
web	56.32	50.32	55.70	51.43	51.30	51.40	51.69	50.74	56.49	50.17
cleveland	54.03	51.61	54.96	54.03	52.35	51.54	51.54	51.54	52.68	53.31
glass	67.54	67.54	67.54	62.25	66.87	66.31	66.02	67.54	67.44	67.44
heart	75.56	74.74	76.74	77.11	77.11	74.15	74.15	74.15	75.48	76.37
olitos	66.67	60.67	60.50	63.00	62.00	61.83	60.67	60.67	57.67	65.67
water2	82.56	83.49	83.44	81.95	81.74	82.10	82.41	83.69	81.18	81.44
water3	82.67	80.92	81.28	81.13	79.59	81.08	79.79	80.62	76.82	77.95
wine	93.82	95.39	93.82	79.54	87.29	91.39	95.05	95.27	88.73	90.86

TABLE II: Classification results (%) using the J48 classifier learner

Dataset	Unred.	GHC	FRFG τ =						GA	PSO
			0.0	0.2	0.4	0.6	0.8	0.9		
MIAS	63.74	60.94	57.09	63.10	60.84	61.74	61.19	60.26	64.41	53.34
DDSM	5278	49.22	48.88	51.14	50.21	49.65	47.77	48.73	51.79	50.69
web	54.74	49.68	55.94	51.40	51.57	50.22	52.66	51.96	61.45	50.70
cleveland	54.23	54.48	55.22	53.22	54.41	54.48	54.48	54.48	54.02	54.09
glass	67.17	67.17	67.17	60.56	66.68	65.05	64.95	67.17	65.25	65.25
heart	72.96	74.15	74.15	74.44	74.96	73.93	73.93	73.93	72.30	73.85
olitos	68.50	62.83	60.00	61.67	59.00	59.50	59.67	59.67	59.33	61.17
water2	82.15	83.28	82.87	82.15	82.05	82.21	82.97	83.69	82.00	81.90
water3	82.72	81.23	82.56	81.18	80.36	81.28	80.87	81.74	78.82	78.00
wine	93.54	91.46	92.69	76.61	86.72	90.33	93.25	93.38	86.60	90.41

TABLE III: Classification results (%) using the JRip classifier learner

Dataset	Unred.	GHC	FRFG τ =						GA	PSO
			0.0	0.2	0.4	0.6	0.8	0.9		
MIAS	69.57	63.29	58.72	63.38	62.64	63.16	61.38	58.61	65.40	53.48
DDSM	51.55	45.85	45.34	46.97	47.63	45.39	45.85	46.00	52.13	46.71
web	37.98	44.11	39.20	48.83	45.08	45.32	42.77	41.07	46.72	36.65
cleveland	56.98	52.96	56.91	50.79	54.77	52.96	52.96	52.96	53.89	53.83
glass	69.24	69.24	69.24	59.87	63.28	68.23	68.52	69.24	68.51	68.51
heart	80.96	78.15	81.11	75.85	79.85	77.56	77.56	77.56	78.15	76.96
olitos	81.00	65.67	65.67	66.33	67.67	65.67	66.83	66.83	66.50	72.33
water2	85.33	84.56	87.08	84.97	82.21	83.49	84.77	85.33	78.26	80.10
water3	82.97	81.23	86.36	80.92	81.54	82.51	80.36	80.92	77.44	77.23
wine	95.97	96.42	96.96	73.75	90.21	92.59	95.15	95.05	91.82	94.71

TABLE IV: Classification results (%) using the IBk (kNN) classifier learner (k=3)

same also for PSO, but the FRFG approach actually offers results which are statistically better than PSO for five of the datasets, most notably *wine* and *MIAS*. When considering the unreduced data, the classification results are statistically equivalent, indicating that good features are selected using the FRFG approach.

Considering the average subset size as shown in table V, the FRFG approach returns a range of results which seem to be similar to, or better than those of GHC. varying the value of τ generally tends to result in larger or smaller average subset size, depending on the dataset. For this comparison, the results for τ=0 are ignored as it is essentially a ranking of features, followed by the linear addition to the reduct candidate as they appear in the ranked list. For the *olitos, heart, water2* and *water3* datasets in particular, the average subset size does not seem to change significantly when $\tau \geq 0.6$. In terms of GA and PSO, the FRFG approach demonstrates a significant improvement in performance for the larger datasets: *MIAS, DDSM* and *web*. For the smaller datasets, the pattern seems to be that of equivalent or better performance (disregarding any particular value of τ).

One of the primary motivations behind the development of the FRFG approach was that of a reduction in computational overhead. Many of the fuzzy-rough metrics suffer in this respect when applied to large datasets and using existing search methods. It is clear from table VI, that FRFG has

Dataset	GHC	FRFG τ =						GA	PSO
		0.0	0.2	0.4	0.6	0.8	0.9		
MIAS	6.08	19.02	4.50	6.40	6.28	6.24	6.22	9.0	7.70
DDSM	7.12	34.26	4.94	7.44	7.32	7.10	7.16	10.96	9.56
web	19.02	496.40	28.42	22.00	19.64	19.40	19.06	186.00	141.20
cleveland	7.64	12.08	5.52	6.40	6.28	7.64	6.22	9.0	7.70
glass	9.00	9.00	3.16	5.02	8.00	8.12	9.00	8.36	8.36
heart	7.06	11.00	5.24	8.06	7.06	7.06	7.06	7.00	7.38
olitos	5.00	6.38	5.52	5.04	5.00	5.00	5.00	5.24	5.00
water2	6.00	6.98	6.86	6.10	6.04	6.00	6.00	6.96	6.44
water3	6.08	7.80	6.76	6.16	6.04	6.00	6.00	7.00	6.50
wine	5.00	5.40	1.80	4.88	4.94	4.98	5.00	4.70	4.92

TABLE V: Average subset sizes

Dataset	GHC	FRFG τ =						GA	PSO
		0.0	0.2	0.4	0.6	0.8	0.9		
MIAS	12.04	0.96	0.57	0.8	1.09	1.61	2.76	3.11	22.60
DDSM	110.44	4.17	2.32	4.32	8.26	12.85	25.09	23.94	173.93
web	98.42	11.60	13.29	18.53	37.21	67.24	81.57	3.51	24.07
cleveland	0.39	0.13	0.14	0.40	0.45	0.43	0.45	16.20	3.83
glass	0.14	0.07	0.05	0.08	0.15	0.15	0.19	1.55	1.08
heart	0.30	0.11	0.11	0.31	0.36	0.34	0.35	14.48	3.46
olitos	0.11	0.05	0.06	0.09	0.12	0.14	0.15	2.36	1.26
water2	2.16	0.18	0.40	0.69	0.98	1.48	1.87	20.14	19.71
water3	2.17	0.20	0.41	0.75	1.03	1.5	1.87	19.57	17.25
wine	0.11	0.04	0.04	0.08	0.12	0.13	0.16	7.55	1.29

TABLE VI: Average execution times per fold (sec.)

much to offer in addressing this problem particularly when the larger datasets in this work are considered. Ostensibly, it would appear that GA-based search performs well for the *web* dataset, however if the corresponding results in table V are considered, it can be seen that the average subset size is over 6.5 times that of the worst case for FRFG. The ability of FRFG to return more compact or similar sized subsets for large data whilst doing so in a much reduced execution time are encouraging. It seems that whilst FRFG offers some advantage for the smaller datasets, this varies with respect to the value of τ. This is most likely related to the process of formation of the groups. For datasets of smaller dimensionality, it may not be realistic to form reasonable groups based on higher values of τ as there may be lower levels of overall redundancy.

V. Conclusion

This paper has presented a new approach to feature selection based on feature grouping in order to reduce computational effort. The approach is a modified hill-climber based around the grouping of similar features together using a measure of relatedness, prior to the final selection phase. The internal ranking of these groups is then used in order to guide the search and selection of representative features from each group. For the work described here a fuzzy-rough subset evaluation is employed as a metric in order to determine the goodness of the subset candidate. The experimental evaluation has demonstrated that the approach is particularly useful for larger datasets and that in terms of classification performance, it is at the very least comparable to GHC. When compared to the nature inspired/stochastic approaches (GA and PSO), the proposed approach easily outperforms these in terms of time taken and subset quality.

The approach and ideas described in this paper offer some new directions for further development. In particular, (and as mentioned previously) the FRFG algorithm is a general approach, and it is not limited to the use of the fuzzy-rough set subset evaluator and indeed any metric can be used for this purpose. As such, it would be interesting to investigate the advantages for other metrics, particularly those which perform well but may not scale-up for larger datasets. One of the other aspects that may provide some additional potential for the approach is an in-depth investigation of the effects of the choice of value for the parameter τ. This may provide some insight into how the value can be selected automatically or indeed derived from the data.

Another important factor is how groups are formed; in the present approach, the sample correlation is used as the basis for group membership. Although this means that the number of groups is initially the same as the number of features, the impact of this is reduced by the use of `moreAvoids` and the appropriate choice of parameter value for τ. This may still pose a problem for very large datasets, however, so an alternative feature clustering scheme could be adopted in order to ensure quick clustering and small group sizes. One such

clustering mechanism is presented in [11], which employs a rough set discernibility-based attribute similarity measure for identifying interchangeable groups of attributes. This could be extended to fuzzy-rough discernibility and utilised in the present work, resulting in a true fuzzy-rough approach to group-based feature selection.

Although the experimental evaluation in this paper features at least three large datasets, the application of FRFG to data (particularly real-world data) in the order of thousands of features and objects would also form the basis for a series of future investigations.

ACKNOWLEDGMENT

Neil Mac Parthaláin would like to acknowledge the financial support for this research through NISCHR (*National Institute for Social Care and Health Research)* Wales, Grant reference: RFS-12-37.

REFERENCES

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

[2] C. Cornelis, R. Jensen, G. Hurtado Martín, D. Ślęzak, Attribute Selection with Fuzzy Decision Reducts, Information Sciences, vol. 180, no. 2, pp. 209–224, 2010.

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

[4] R. Diao and Q. Shen. Feature selection with harmony search, IEEE Trans. Syst., Man, Cybernetics Part B, vol. 42, No.6, pp. 1509–1523, 2012.

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

[6] A. Frank and A. Asuncion. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[7] I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[8] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P.J. Kegelmeyer. The Digital Database for Screening Mammography. Proceedings of the Interna- tional Workshop on Digital Mammography, pp. 212–218, 2000.

[9] Q. Hu, D. Yu, and Z. Xie. Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters, vol. 27, no.5, pp. 414–423, 2006.

[10] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Perez, E.R.E. Denton, R. Zwiggelaar. A Novel Breast Tissue Density Classification Methodology. IEEE Transactions on Information Technology in Biomedicine, vol. 12, no. 1, pp. 55–65, 2008.

[11] A. Janusz and D. Ślęzak. Utilization of attribute clustering methods for scalable computation of reducts from high-dimensional data, 2012 Federated Conference on Computer Science and Information Systems (FedCSIS), pp.295–302, 2012.

[12] R. Jensen and Q. Shen. New Approaches to Fuzzy-Rough Feature Selection, IEEE Transactions on Fuzzy Systems, vol. 17, no. 4, pp. 824–838, 2009.

[13] R. Jensen, A. Tuson, and Q. Shen. Finding rough and fuzzy-rough set reducts with SAT, Information Sciences, vol. 255, pp. 100–120, 2014.

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

[15] L. Polkowski. Rough Sets: Mathematical Foundations, Advances in Soft Computing, Physica Verlag, Heidelberg, Germany, 2002.

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

[17] 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.

[18] S. Stawicki and S. Widz. Decision bireducts and approximate decision reducts: Comparison of two approaches to attribute subset ensemble construction, Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 331–338, 2012.

[19] J Suckling, J. Partner, D.R. Dance, S.M. Astley, I. Hutt, C.R.M. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S.L. Kok, P. Taylor, D. Betal, and J. Savage. The Mammographic Image Analysis Society digital mammogram database. International Workshop on Digital Mammography, pp. 211–21, 1994.

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

[21] L.A. Zadeh. Fuzzy sets, Information and Control, vol. 8, no. 3, pp. 338–353, 1965.