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

Heuristic Search for Fuzzy-Rough Bireducts and its Use in Classifier Ensembles

Ren Diao, Neil Mac Parthaláin, Richard Jensen and Qiang Shen
Department of Computer Science
Institute of Mathematics, Physics and Computer Science
Aberystwyth University, UK
Email: {rrd09, ncm, rkj, qqs}@aber.ac.uk

***Abstract*—Rough set theory has proven to be a useful mathematical basis for developing automated computational approaches which are able to deal with and utilise imperfect knowledge. Fuzzy-rough set theory is an extension to rough set theory and enhances the ability to model uncertainty and vagueness more effectively. There have been many developments in this area which offer robust methods for feature selection or instance selection. However, these are often carried out in isolation rather than considering both types of selection simultaneously. For this purpose, the notion of a bireduct has been proposed recently but the task of finding bireducts of high quality remains a significant challenge. This paper presents a heuristic strategy for the identification of fuzzy-rough bireducts, which is based on a music-inspired global optimisation algorithm called harmony search. The concept of ϵ-bireducts is employed in this approach for the evaluation and improvisation of the candidate solutions. The stochastically-selected bireducts are also utilised to construct classifier ensembles. The presented technique is experimentally evaluated using a number of real-valued benchmark data sets.**

I. Introduction

Feature selection (FS) is becoming an increasingly necessary step for today's ever-growing data sets. Its primary aim is to discover a minimal feature subset for a problem domain, whilst retaining a suitably high accuracy in representing the original data [1]. When analysing data with very large numbers of features [2], it is difficult to identify and extract patterns or rules due to the high inter-dependency amongst individual features, or the combined behaviour of groups of features. Techniques that perform tasks such as object recognition [3], data classification [4], and systems monitoring [5] can benefit significantly, when the noisy, irrelevant, redundant or misleading features are removed [6].

Data is currently being collected and archived at a staggering pace in almost every field imaginable. In addition, data sets themselves grow larger in terms of both the number of measurements (features), and the number of data instances. This continued growth places high demand on resources for both the storage and maintenance of data. In addition to FS, approaches developed for the purpose of instance selection [7] (IS) are also desirable, as they may help to considerably reduce the volume of the data, whilst simultaneously removing those misleading or noisy training instances. Therefore, both FS and IS are techniques which reduce the dimensionality of data and help to improve any models that are subsequently learned from that data.

Rough set theory (RST) [8] has attracted great interest amongst researchers in recent years. Its popularity stems from a multitude of appealing theoretical aspects. Indeed, the focus of RST on grouping information entities into "granules" in terms of a certain form of relatedness offers a certain universal intuitive appeal. In addition, it has other desirable attributes; for example, no tunable parameters are required, thus eliminating the need for (possibly erroneous) subjective human intervention. It also finds a minimal knowledge representation. One of the problems for RST however, is that it is constrained to crisp or discrete-valued data, and its inability to deal with real-valued data has led to the development of fuzzy-rough sets [9].

Much work has been carried out in the area of FS using both rough and fuzzy-rough sets [10]. The vast majority of such work focuses on the use of decision reducts. These are subsets of features which fully characterise the knowledge in the data. Rough set bireducts [11] are a newly proposed concept that further extends the idea of a decision reduct. Bireducts place emphasis on both the subset of features, which describe decisions, *and* the subset of data instances for which such a description is valid. A bireduct therefore essentially offers a representation for a sub-table of the data characterised by subsets of both features and data instances. The properties of rough set bireducts have also been exploited and extended to fuzzy-rough sets [12], such that they can be applied to real-valued data. A boolean representation in conjunctive normal form (CNF) has also been adopted, in order to transform the task of simultaneous instance and feature selection into a constraint satisfaction problem [12]. Whist the work in [11] and [12] offer much in terms of extending the underlying rough and fuzzy-rough concepts respectively, defining the optimality of any given bireduct is a difficult and challenging task.

In an attempt to address this challenge, this paper presents a heuristic search strategy, which provides an alternative means for the identification of potentially optimal fuzzy-rough bireducts. A recently developed FS search algorithm based on harmony search [13] is modified, and the notion of an ε-bireduct [11] is exploited, in order to better identify the most suitable bireducts (judged by metrics such as object coverage and feature subset size). As a stochastic approach, the proposed algorithm is capable of discovering multiple bireducts of similar quality, which are further utilised to construct classifier ensembles.

The remainder of this paper is structured as follows. Section II summarises the theoretical aspects of fuzzy-rough bireducts. Section III explains the proposed heuristic algorithm aimed at searching for potentially "optimal" candidate solutions. Section IV points out the benefits of employing a bireduct-based classifier ensemble, and outlines its structure and working

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

procedures. Section V presents the experimental results that demonstrate the efficacy of the proposed approach. Section VI concludes the paper and identifies a number of areas for future development.

II. Theoretical Background

In the context of FS, an information system is a tuple $\langle \mathbb{U}, \mathbb{A} \rangle$, where $\mathbb{U} = \{x_1, \cdots, x_{|\mathbb{U}|}\}$ is a non-empty set of finite objects (commonly referred to as the universe of discourse); and $\mathbb{A} = \{a_1, \cdots, a_{|\mathbb{A}|}\}$ is a non-empty, finite set of features such that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that feature a may take, and $|\mathbb{A}|$ denotes the cardinality of the set $\mathbb{A}$. For decision systems, there exists a set of decision features Z that jointly form an injunction to the input features $\mathbb{A}$, which may be either discrete- or real-valued. In this paper, for simplicity, a single decision feature d is considered, i.e., $Z = \{d\}$.

A. Rough and Fuzzy-Rough Sets

Rough set theory (RST) has been successfully used for the task of FS in order to discover data dependencies and to reduce the number of features contained in a data set [9]. Given a data set with discrete feature values, RST can find a subset (termed a *reduct*) of the original features that are the most informative; all other features can be removed from the data set with minimal information loss. Traditional RST only works on discrete, crisp-valued domains, however in practice, the values of features are usually real-valued. It is not possible in the theory to say whether two different feature values are similar, and to what extent they may be equivalent. For example, two close values may only differ as a result of noise, but in the standard RST-based approach they are considered to be as different as two values of a different order of magnitude. Data set discretisation must therefore take place before reduction methods based on crisp rough sets can be applied. This is often still inadequate, however, as the degrees of membership of values to discretised values are not considered and thus may result in information loss. In order to combat this, extensions of RST such as fuzzy-rough sets [10] have been developed.

A fuzzy-rough set is defined by two fuzzy sets, a fuzzy lower and a fuzzy upper approximation, obtained by extending the corresponding crisp rough set notions. In the crisp case, elements either belong to the lower approximation with absolute certainty or not at all. In the fuzzy-rough case, elements may have a membership in the range [0,1], allowing greater flexibility in handling uncertainty. Fuzzy-rough FS [6] (FRFS) extends the ideas of fuzzy-rough sets to perform FS, with the following definitions:

$$\mu_{\underline{R_B}X}(x_i) = \inf_{x_j \in \mathbb{U}} \mathcal{I}(\mu_{R_B}(x_i, x_j), \mu_X(x_j)) \tag{1}$$

$$\mu_{\overline{R_B}X}(x_i) = \sup_{x_j \in \mathbb{U}} \mathcal{T}(\mu_{R_B}(x_i, x_j), \mu_X(x_j)) \tag{2}$$

where X is the fuzzy concept being approximated, $\mathcal{I}$ is a fuzzy implicator, $\mathcal{T}$ is a t-norm, and R_B is the fuzzy similarity relation induced by the subset of features B, and $x_i, x_j \in X$ are two arbitrary objects in X. In particular,

$$\mu_{R_B}(x_i, x_j) = \mathcal{T}_{a \in B}\{\mu_{R_a}(x_i, x_j)\} \tag{3}$$

where $\mu_{R_a}(x_i, x_j)$ is the degree to which objects x_i and x_j are similar for feature $a \in \mathbb{A}$. Many similarity relations can be constructed for this purpose, for example:

$$\mu_{R_a}(x_i, x_j) = 1 - \frac{|a(x_i) - a(x_j)|}{a_{\max} - a_{\min}} \tag{4}$$

$$\mu_{R_a}(x_i, x_j) = \exp(-\frac{(a(x_i) - a(x_j))^2}{2\sigma_a^2}) \tag{5}$$

$$\begin{aligned}\mu_{R_a}(x_i, x_j) = \max(\min(&\frac{a(x_j) - (a(x_i) - \sigma_a)}{a(x_i) - (a(x_i) - \sigma_a)}, \\ &\frac{(a(x_i) + \sigma_a) - a(x_j)}{(a(x_i) + \sigma_a) - a(x_i)}), 0)\end{aligned} \tag{6}$$

where σ_a^2 is the variance of feature a, and $a(x_i)$ is the value of feature a for object x_i. The choices for $\mathcal{I}$, $\mathcal{T}$, and the fuzzy similarity relation have great influence upon the resultant fuzzy partitions.

B. Fuzzy Discernibility Matrices

A number of recent developments in fuzzy-rough set-based FS focus on the use of fuzzy discernibility matrices and functions [14]. A fuzzy discernibility matrix may be represented by a set of fuzzy clauses $\mathbb{C}$, where each of the fuzzy clauses $C_{ij} \in$ is a fuzzy set, to which every feature $a \in \mathbb{A}$ belongs to a certain degree $\mu_{C_{ij}}(a)$, determined by the fuzzy discernibility measure:

$$\mu_{C_{ij}}(a) = \mathcal{N}(\mu_{R_a}(x_i, x_j)) \tag{7}$$

where $\mathcal{N}$ denotes the fuzzy implementation of the negation operator ($\neg$).

A fuzzy discernibility function $f_{\mathbb{C}}$ can then be defined on the basis of fuzzy clauses, which is best represented in conjunctive normal form (CNF):

$$f_{\mathbb{C}}(B) = f_{\mathbb{C}}(a_1^*, \cdots, a_{|\mathbb{A}|}^*) = \wedge\{\vee C_{ij}^* \mid C_{ij} \in \mathbb{C}\} \tag{8}$$

where $1 \leq i < j \leq |\mathbb{U}|$, and the truth assignment a_i^* of a given feature $a_i \in \mathbb{A}$ is:

$$a_i^* = \begin{cases} \texttt{true} & , \ a_i \in B \\ \texttt{false} & , \ \text{otherwise} \end{cases} \tag{9}$$

This function returns values in $[0, 1]$, and reflects the extent to which the function is satisfied for a given assignment of truth values to the variables $\{a_1, \cdots, a_{|\mathbb{A}|}\}$.

For decision systems, only those clauses with "different" decision values are considered, and therefore, the fuzzy discernibility function is modified in order to reflect the following logical operations:

$$f_{\mathbb{C}}(a_1^*, \cdots, a_{|\mathbb{A}|}^*) = \wedge\{\vee\{C_{ij}^*\} \rightarrow \neg(d(x_i) = d(x_j))\} \tag{10}$$

so that the extent to which decision values differ may affect the overall satisfiability of the clauses.

Note that the degree of satisfaction for a clause C_{ij} regarding a given feature subset B is defined as:

$$\text{SAT}_B(C_{ij}) = \mathcal{S}_{a \in B}\{\mu_{C_{ij}(a)}\} \tag{11}$$

for a t-conorm $\mathcal{S}$, which is used to determine the clauses satisfiable by the selected features. Unlike traditional (crisp)

propositional satisfiability, a fuzzy clause may be satisfied to a certain degree depending on what extent to which variables have been assigned the value `true`. Obviously, the maximum satisfiability degree of a given clause $\text{SAT}_{\max}(C_{ij})$ can be achieved by assigning all of its involved features to `true`. A minimal fuzzy-rough reduct R can therefore be defined as a (minimal) truth assignment to variables (features), such that each clause $C_{ij} \in \mathbb{C}$ is satisfied to its maximum extent:

$$R = \arg\min_{B \subseteq \mathbb{A}} |B|, f_{\mathbb{C}}(B) = f_{\mathbb{C}}(\mathbb{A}) \tag{12}$$

C. Fuzzy-Rough Discernibility-Based Bireducts

The notion of a fuzzy-rough bireduct is inspired by the idea of a rough set bireduct [11], which is originated from a similar, yet non-equivalent approach termed approximate reducts [15]. The definition of such a concept focuses on a subset of features that jointly describe the decision feature d, and a subset of instances $Y \subseteq \mathbb{U}$ for which such a description is valid [12].

Definition 1: Let $\langle \mathbb{U}, \mathbb{A} \rangle$ be an information system. A tuple (B, Y), where $B \subseteq \mathbb{A}$ and $Y \subseteq \mathbb{U}$ is an information bireduct, iff B discerns all pairs of objects in Y, $\forall x_i, x_j \in Y, \exists a \in B | a(x_i) \neq a(x_j)$, and:

- There is no proper subset $B' \subset B$ such that B' discerns all pairs in Y.
- There is no proper superset $Y' \supset Y$ such that B discerns all pairs in Y'.

Definition 2: Let $\langle \mathbb{U}, \mathbb{A} \cup \{d\} \rangle$ be a decision system, a tuple (B, Y), where $B \subseteq \mathbb{A}$ and $Y \subseteq \mathbb{U}$ is a decision bireduct, iff B discerns all pairs of objects in Y, where $d(x_i) \neq d(x_j)$, and:

- There is no proper subset $B' \subset B$ such that B' discerns all pairs $x_i, x_j \in Y$, where $d(x_i) \neq d(x_j)$.
- There is no proper superset $Y' \supset Y$ such that B discerns all pairs $x_i, x_j \in Y'$, where $d(x_i) \neq d(x_j)$.

The properties of a bireduct ensure that the feature subset contained by the bireduct is minimal, and that the coverage of the data instances is maximal. By employing the CNF representation introduced earlier in Section II-B, a revised fuzzy discernibility function [12] may be established in order to facilitate the bireduct scenario:

$$f_{\mathbb{C}}(B, Y) = f_{\mathbb{C}}(a_1^*, \cdots, a_{|\mathbb{A}|}^*, x_1^*, \cdots, x_{|\mathbb{U}|}^*) = \wedge \{x_i^* \vee x_j^* \vee C_{ij}^*\}$$

where

$$x_i^* = \begin{cases} \texttt{true} & , \ x_i \notin Y \\ \texttt{false} & , \ \text{otherwise} \end{cases} \tag{13}$$

A clause C_{ij} may now be satisfied if C_{ij}^* is maximally satisfied, or if either x_i or x_j is "selected" (which means that the corresponding training instance will be excluded from the set of covered objects Y). Consequently, these "selected" objects form the set of outliers $O = \mathbb{U} \setminus Y$.

III. Harmony Search for Fuzzy-Rough Bireducts

Many heuristic (and stochastic) search strategies have been exploited for the task of FS, in an effort to identify compact and good quality feature subsets without resorting to exhaustive search. Inspired by natural phenomena or patterns of social behaviour, a good number of such methods have shown promising results in dealing with complex problem scenarios. Harmony search (HS) [13] for instance, is a music-inspired technique that is particularly effective in consolidating the size of the emerging feature subsets. This section explains the modifications made to the original HS-based FS algorithm (HSFS), in order to facilitate the recognition and optimisation of fuzzy-rough bireducts.

A. Mapping of Key Notions

The key notions of HS are musicians, notes, harmony, and harmony memory. For conventional optimisation problems, the musicians $P = \{p^i \mid i = 1, \cdots, |P|\}$ represent the variables of the cost function being optimised, their values are referred to as notes. A harmony H is a candidate solution vector containing the values for each variable, where a collection of good quality harmonies are stored in the harmony memory $\mathbb{H} = \{H^j \mid j = 1, \cdots, |\mathbb{H}|\}$. Note that P, H, and $\mathbb{H}$ are ordered lists rather than sets. In particular, $H_i^j, i = 1, \cdots, |P|, j = 1, \cdots, |\mathbb{H}|$, denotes the value selected by the i^{th} musician for the j^{th} harmony. The harmony memory $\mathbb{H}$ can be concretely represented by a two dimensional matrix of a rank $|\mathbb{H}| \times |P|$. Without loss of generality, the number of rows (harmonies) $|\mathbb{H}|$ is a predefined parameter, and each column is dedicated to one musician, which provides a pool of playable notes (referred to hereafter as the note domain $\aleph^i$ of a musician p^i, $\aleph^i = \bigcup_{j=1}^{|\mathbb{H}|} H_i^j$) for future improvisations.

When applied to FS, as shown in Table I, a musician is best described as an independent expert or a "single-feature selector", and the available features translate to notes. Each musician may vote for one feature to be included in the emerging harmony, which is the combination of votes from all such single-feature selectors, indicating the features to be selected. The available features $\{a_1, \cdots, a_{|\mathbb{A}|}\}$ form the pool of playable notes, which is shared by all of the musicians. Multiple musicians are allowed to choose the same feature, or they may opt to choose none at all. The fitness function fit(H) becomes a feature subset evaluator that analyses and merits each of the new feature subsets found during the search process.

TABLE I. Concept mapping from HS to FS

HS	Optimisation	FS
Musician	Variable	Single-Feature Selector
Musical Note	Variable Value	Feature
Harmony	Solution Vector	Feature Subset
Harmony Memory	Solution Storage	Feature Subset Storage
Harmony Evaluation	Fitness Function	Feature Subset Evaluation
Optimal Harmony	Optimal Solution	Optimal Feature Subset

The proposed HSFS algorithm for fuzzy-rough bireducts (referred to as HSFS_{BR} hereafter) uses 4 parameters: the size of the harmony memory $|\mathbb{H}|$, the number of "single-feature selectors" (musicians) $|P|$, a harmony memory considering rate δ, and the maximum number of iterations $g_{\max}$. The parameter $\delta, 0 \leq \delta \leq 1$ controls the rate at which a selector p^i can randomly choose a feature from all available features $\mathbb{A}$ (instead of within its own note domain $\aleph^i$). For example, if δ has a value of 0.85, the musicians have a 15% chance to explore alternative features, which may potentially lead to better quality feature subsets. For the remaining 85% of time, the musicians focus on improving existing solutions from values within their

TABLE II. HARMONY ENCODED FEATURE SUBSETS

	p^1	p^2	p^3	p^4	p^5	p^6	Represented Subset B
H^1	a_2	a_1	a_3	a_4	a_7	a_{10}	$\{a_1, a_2, a_3, a_4, a_7, a_{10}\}$
H^2	a_2	a_2	a_2	a_3	a_{13}	$-$	$\{a_2, a_3, a_{13}\}$
H^3	a_2	$-$	a_2	$a_3 \rightarrow a_6$	a_{13}	a_4	$\{a_2, a_4, a_6, a_{13}\}$

respective note domains. HSFS relies on stochastic mechanisms such as this in order to escape from local minimal solutions.

Table II depicts the following three example harmonies. H^1 denotes a subset of 6 distinctive features: $B_{H^1} = \{a_1, a_2, a_3, a_4, a_7, a_{10}\}$. H^2 shows a duplication of choices from the first three musicians, and a discarded note (represented by $-$) by p^6, representing a reduced subset $B_{H^2} = \{a_2, a_3, a_{13}\}$. H^3 signifies the feature subset $B_{H^3} = \{a_2, a_6, a_4, a_{13}\}$, where $a_3 \rightarrow a_6$ indicates that p^4 originally nominated a_3, but it is forced to change its choice to a_6 due to δ activation. For simplicity, the explicit encoding/decoding process between a given harmony H^j and its associate feature subset B_{H^j} is omitted in the following descriptions.

B. Evaluation of Fitness and ε-Bireducts

Fitness functions play an important role in any search algorithm, as they guide the search process towards better candidate solutions. The optimality of a given bireduct (B, Y) may be interpreted from a number of different perspectives, such as the compactness of the feature subset: $|B|$, the coverage of the selected features: $|Y|$, and the balance between the number of features and objects involved. Furthermore, a given feature subset B may form multiple bireducts, since the CNF representation shown in Eqn. II-C may be satisfied using many alternative combinations of objects. The associated family of subsets (of covered objects) $\mathbb{Y}_B = \{Y \mid f_{\mathbb{C}}(B, Y) = f_{\mathbb{C}}(A, \mathbb{U}), B \subseteq \mathbb{A}, Y \subseteq \mathbb{U}\}$ may also vary in size.

In order to better judge the quality of a given bireduct, the notion of an ε-bireduct has been introduced [11]:

$$|Y| \geq (1 - \varepsilon)|\mathbb{U}|, 0 \leq \varepsilon < 1 \tag{14}$$

which attempts to impose a constraint over the emerging bireducts. Different settings of ε have a direct impact upon the size of Y, which in turn influence the size of the emerging feature subsets. Analogies of the ε-bireduct concept have been studied for frequent itemsets and patterns [16]. Intuitively, the fewer the objects to be covered (higher ε), the fewer the number of features is generally required. A ε-bireduct collapses to a standard reduct when $\varepsilon = 0$, since the definition will require that all the training objects are fully covered by the selected features.

In order to identify the "minimal" ε-bireducts for a given value of ε, i.e., a bireduct (B, Y) with a low cardinality of B, in this paper, the fitness of a given harmony H (and its associated feature subset B_H) is calculated as follows:

$$\text{fit}(H) = \begin{cases} \text{cov}(B_H), & \text{cov}(B_H) \leq 1 - \varepsilon \\ 2 - 2\varepsilon - \text{cov}(B_H), & \text{cov}(B_H) > 1 - \varepsilon \end{cases} \tag{15}$$

where the highest achievable fitness value is $1 - \varepsilon$. Here $\text{cov}(B_H)$ represents the maximum number of objects coverable using feature subset B_H:

$$\text{cov}(B_H) = \max_{Y \in \mathbb{Y}_{B_H}} \left(\frac{|Y|}{|\mathbb{A}|}\right) \tag{16}$$

The feature subsets are also compared on the basis of size, since the ultimate goal is to identify a compact subset of features that is able to fully describe a sufficient amount of objects, and thus, the HSFS_{BR} algorithm is bi-objective by nature as with the underlying problem itself.

C. Iteration Steps of $HSFS_{BR}$

The overall operation of the proposed HSFS_{BR} algorithm is illustrated in Fig. 1 and outlined in Algorithm 1. The following explains it briefly.

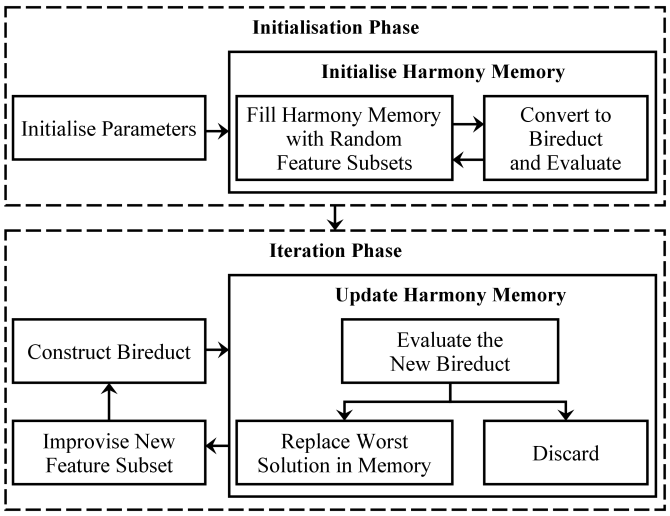

Fig. 1. Work flow of HSFS_{BR}

1) Initialise Harmony Memory: Set the initial values for the parameters $|\mathbb{H}|$, $|P|$, δ, and $g_{\max}$ as with the application of conventional FS. A harmony memory containing $|\mathbb{H}|$ randomly generated feature subsets is then initialised. The fuzzy-rough bireducts associated with these feature subsets are then identified, and fitness values calculated following Eqns. 15 and 16. This initialisation procedure also ensures that each of the single-feature selectors has a note domain $\aleph$ of $|\mathbb{H}|$ features, which may include identical choices, or nulls/discards ($-$).

2) Improvise New Subset: Each p^j in P nominates a feature $a \in \aleph^j$ and all such nominated features form a new harmony H^{new}. The evaluation score of the corresponding new feature subset $B_{H^{\text{new}}}$, decoded by following a scheme that generalise what is illustrated in Table II, can then be computed according to Eqns. 15 and 16.

3) Update Subset Storage: If the newly obtained subset achieves a higher evaluation score than that of the worst subset in the harmony memory, or if it has an equal evaluation but is of a smaller size, then this new feature subset replaces the existing worst feature subset. Otherwise, it is discarded.

4) Iterate: The improvisation-update process repeats until the maximum number of iterations is reached. In the end, the best harmony in the harmony memory $\dot{H} = \arg\max_{H \in \mathbb{H}} \text{fit}(B_H)$ and its associated ε-bireduct $(B_{\dot{H}}, Y)$ are returned as the final search output.

Algorithm 1: HSFS_{BR} Algorithm

$p^i \in P$, $i = 1$ **to** $|P|$, group of musicians
$H^j \in \mathbb{H}$, $j = 1$ **to** $|\mathbb{H}|$, harmony memory
$\aleph_i = \bigcup_{j=1}^{|\mathbb{H}|} H_i^j$, note domain of p^i
δ, harmony memory considering rate
$\mathbb{C}$, fuzzy clauses (fuzzy discernibility matrix)
for $g = 1$ **to** $g_{\max}$ **do**
  $H^{\text{new}} = \emptyset$
  for $i = 1$ **to** $|P|$ **do**
    $r_\delta =$ a random real number, $0 \leq r_\delta \leq 1$
    if $r_\delta < \delta$ **then**
      $a_r =$ a random feature, $a_r \in \mathbb{A}$
      $H^{\text{new}} = H^{\text{new}} \cup \{a_r\}$
    else
      $r =$ a random integer, $1 \leq r \leq |\mathbb{H}|$
      $H^{\text{new}} = H^{\text{new}} \cup \{\aleph_{ir}\}$
  for $\forall C_{ij} \in \mathbb{C}$ **do**
    if $SAT_{B_{H^{new}}}(C_{ij}) = SAT_{\max}(C_{ij})$ **then**
      $\mathbb{C} = \mathbb{C} \setminus C_{ij}$
  Identify the outliers O that satisfies the remaining $\mathbb{C}$
  Form bireduct $(B_{H^{\text{new}}}, \mathbb{U} \setminus O)$
  if $fit(H^{new}, \varepsilon) \geq \min_{H \in \mathbb{H}} fit(H, \varepsilon)$ **then**
    $\mathbb{H} = \mathbb{H} \cup \{H^{\text{new}}\}$
    $\mathbb{H} = \mathbb{H} \setminus \{\arg\min_{H \in \mathbb{H}} \text{fit}(H, \varepsilon)\}$
return best ε-bireduct in $\mathbb{H}$

D. Worked Example

An example data set shown in Table III with one decision feature $Z = \{d\}$ is employed in order to illustrate the key operations of HSFS_{BR}. Due to space, the calculation of the initial fuzzy similarity measures $R_a(x_i, x_j), x_i, x_j \in \mathbb{U}, a \in \mathbb{A}$ is omitted. Refer to [12] for more details. Another example concerning crisp data and rough set-based bireducts can also be found in [11], which should provide a good initial understanding of the developed bireduct concepts.

TABLE III. EXAMPLE DATA SET

$\mathbb{U}$	a_1	a_2	a_3	d
x_1	-0.4	-0.3	-0.5	no
x_2	-0.4	0.2	-0.1	yes
x_3	-0.3	-0.4	-0.3	no
x_4	0.3	-0.3	0	yes
x_5	0.2	-0.3	0	yes
x_6	0.2	0	0	no

In the present example, a given fuzzy clause, say C_{46}, is represented as $C_{46} = \{x_4^* \vee x_6^* \vee a_1^{0.301} \vee a_2^{1.0} \vee a_3^{0.0}\} \leftarrow d^{1.0}$, where $a_1^{0.301}$ signifies that the degree of membership of feature a_1 in this particular clause is $0.301, \mu_{C_{ij}}(a_1) = 0.301$. This implies that the instances x_4 and x_6 are partially discernible using a_1 with respect to the decision feature ($d^{1.0}$). Due to the inherent properties of the implicators [12], all clauses with $d^{0.0}$ may be removed as they do not influence the returned bireduct. The initial set of clauses $\mathbb{C}$ is therefore:

$$
\begin{aligned}
C_{12}: &\quad \{1^* \vee 2^* \vee a_1^{0.0} \vee a_2^{1.0} \vee a_3^{1.0}\} && \leftarrow d^{1.0}\\
C_{14}: &\quad \{1^* \vee 4^* \vee a_1^{1.0} \vee a_2^{0.0} \vee a_3^{1.0}\} && \leftarrow d^{1.0}\\
C_{15}: &\quad \{1^* \vee 5^* \vee a_1^{1.0} \vee a_2^{0.0} \vee a_3^{1.0}\} && \leftarrow d^{1.0}\\
C_{16}: &\quad \{1^* \vee 6^* \vee a_1^{1.0} \vee a_2^{1.0} \vee a_3^{1.0}\} && \leftarrow d^{1.0}\\
C_{23}: &\quad \{2^* \vee 3^* \vee a_1^{0.301} \vee a_2^{1.0} \vee a_3^{0.964}\} && \leftarrow d^{1.0}\\
C_{26}: &\quad \{2^* \vee 6^* \vee a_1^{1.0} \vee a_2^{0.863} \vee a_3^{0.483}\} && \leftarrow d^{1.0}\\
C_{34}: &\quad \{3^* \vee 4^* \vee a_1^{1.0} \vee a_2^{0.431} \vee a_3^{1.0}\} && \leftarrow d^{1.0}\\
C_{35}: &\quad \{3^* \vee 5^* \vee a_1^{1.0} \vee a_2^{0.431} \vee a_3^{1.0}\} && \leftarrow d^{1.0}\\
C_{46}: &\quad \{4^* \vee 6^* \vee a_1^{0.301} \vee a_2^{0.301} \vee a_3^{0.0}\} && \leftarrow d^{1.0}\\
C_{56}: &\quad \{5^* \vee 6^* \vee a_1^{0.0} \vee a_2^{1.0} \vee a_3^{0.0}\} && \leftarrow d^{1.0}
\end{aligned}
$$

For HSFS_{BR}, the number of musicians is $|P| = |\mathbb{A}| = 3$, and an initial harmony memory is randomly generated, filling the note domains $\aleph_i$ of musicians p^i with randomly selected musical notes (features). A newly improvised harmony H^{new} may be $\{a_3, a_3, -\}$ which represents the feature subset $B_{H^{\text{new}}} = \{a_3\}$. By removing the clauses satisfiable using $a_3^* = \texttt{true}$ alone, the following clauses may be obtained:

$$
\begin{aligned}
C_{23}: &\quad \{2^* \vee 3^* \vee a_1^{0.301} \vee a_2^{1.0} \vee a_3^{0.964}\} && \leftarrow d^{1.0}\\
C_{26}: &\quad \{2^* \vee 6^* \vee a_1^{1.0} \vee a_2^{0.863} \vee a_3^{0.483}\} && \leftarrow d^{1.0}\\
C_{46}: &\quad \{4^* \vee 6^* \vee a_1^{0.301} \vee a_2^{0.301} \vee a_3^{0.0}\} && \leftarrow d^{1.0}\\
C_{56}: &\quad \{5^* \vee 6^* \vee a_1^{0.0} \vee a_2^{1.0} \vee a_3^{0.0}\} && \leftarrow d^{1.0}
\end{aligned}
$$

The maximum object coverage of this feature subset $\{a_3\}$ is either $\{x_1, x_3, x_4, x_5\}$ or $\{x_1, x_2, x_4, x_5\}$, if the objects $\{a_2, a_6\}$ or $\{a_3, a_6\}$ are considered as outliers, respectively, and $\text{cov}(B_{H^{\text{new}}}) = \frac{4}{6}$. If $\varepsilon = 0.4$, then this new harmony will have a fitness evaluation of $2 - 2 \times 0.4 - 0.66 = 0.54$. It replaces the existing worst harmony, if it achieves a higher fitness evaluation (or equal evaluation but more compact in size). The process iterates for $k_{\max}$ numbers of iterations.

IV. CLASSIFIER ENSEMBLE WITH ε-BIREDUCTS

For a given data set of significant complexity, a family $\mathbb{B}$ of quality (while not always equally optimal) feature subsets may be discovered using a stochastic search algorithm. Any such feature subset $B \in \mathbb{B}$ may be used to train a subsequent classifier learner, and a diverse feature subset-based classifier ensemble [17] may be constructed. Ensemble methods commonly achieve better predictive performance than that of a single classifier, as they exploit the uncorrelated errors within the group as a result of their diverse internal models [18].

Following the existing investigations carried out for rough set-based ensemble of bireducts [19], this paper explores the potential of feature subset-based classifier ensembles built on the basis of fuzzy-rough ε-bireducts. The simultaneous selection of both features and training objects [12] are particularly beneficial for the construction of diverse classifiers, since the objects in Y have been selected specifically for the features in B, and are best used to learn from the data using those features. The amount of training objects may be individually configured by specifying ε, which may help to control the space and time complexities of the resultant learned models.

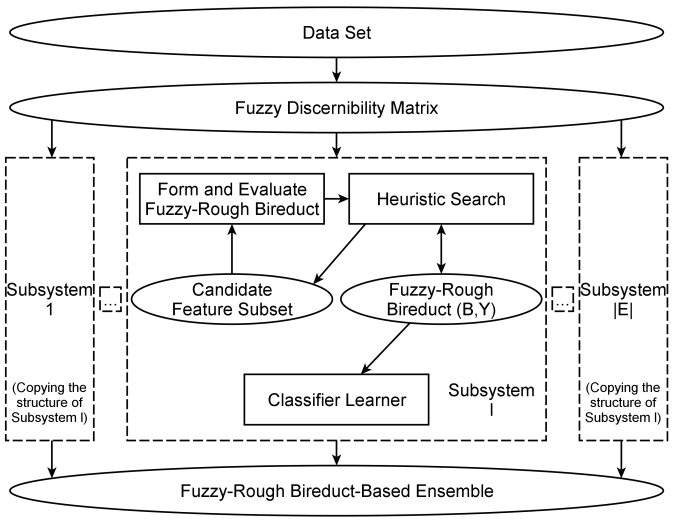

Fig. 2. Generic framework for ε-bireduct-based classifier ensemble

A generic framework for ε-bireduct-based classifier ensemble is presented in Fig. 2. In order to construct an ensemble of classifier $\mathbb{E} = \{E^l \mid l = 1, \cdots, |\mathbb{E}|\}$, a group of bireducts $\{(B^l, Y^l) \mid l = 1, \cdots, |\mathbb{E}|\}$ needs to be identified first. In this figure, each of the dashed blocks of components forms an individual ε-bireduct-based classifier E^l, $l \in \{1, \cdots, |\mathbb{E}|\}$, where the objects in Y^l are employed to train the classification algorithm in conjunction with only the features in B^l.

Although the underlying theoretical notions are different, the work procedure of an ε-bireduct-based classifier ensemble is in principle, similar to that of an ordinary feature subset-based classifier ensemble [17], [20], [21], and that of a rough set-based bireduct ensemble [19]. Therefore, further detailed explanations regarding its operations are omitted due to space. Note that any ensemble aggregation method such as majority vote [22] may be employed to combine the prediction outputs of the base learners.

V. Experimentation

A number of experiments have been carried out in order to demonstrate the capability of the proposed HSFS_{BR} algorithm and the resultant ε-bireduct-based classifier ensemble. In total, nine benchmark data sets taken from the UCI machine learning repository [23] are employed. Information regarding the selected data sets and the parameter settings of HSFS_{BR} employed in the experiments is summarised in Table IV. The classification algorithms adopted in the experiments include two commonly employed techniques: 1) the tree-based C4.5 algorithm [24] which uses entropy to identify the most informative feature at each level, in order to split the training samples according to their respective classes; and 2) the nearest neighbour classifier using vaguely quantified fuzzy-rough sets (VQNN) [25]. Obtaining possibly contrasting views of two different types of base classifier helps to provide a more comprehensive understanding of the qualities of the discovered bireducts.

Stratified 10-fold cross-validation (10-FCV) is employed for result validation. The stratification of the data prior to its division into different folds ensures that each class label has equal representation in all folds, thereby helping to alleviate bias/variance problems [26]. The classifier ensembles are trained using the bireducts identified for each of the cross-validation folds, and then tested for its accuracy using the corresponding test folds. The accuracies of the full (unreduced) data sets are also given for comparison. In this preliminary investigation, ensembles of size 10 ($|\mathbb{E}| = 10$) are employed.

TABLE IV. Parameter Settings and Data Set Information

$\lvert\mathbb{H}\rvert$	$\lvert P\rvert$	δ	$g_{\max}$
10	$\lvert\mathbb{A}\rvert$	0.9	2000
Data set	**Features**	**Objects**	**Classes**
cleveland	14	297	5
ecoli	8	336	8
glass	9	214	6
heart	13	270	2
ionosphere	35	230	2
libras	91	360	15
sonar	61	208	2
water	39	390	3
wine	14	178	3

Tables V to VII show the classification accuracies for the ensembles, which are constructed on the basis of the two base classification algorithms, with respect to three different values of $\varepsilon = 0.1, 0.2, 0.3$. For instance, for Table V, the value of ε is set to 0.1. This means that the bireducts are expected to cover at least 90% of the training objects. The averaged object coverage AVG($\frac{|Y|}{|\mathbb{U}|}$) indicates that the proposed HSFS_{BR} algorithm can indeed identify bireducts according to the specified ε constraints. Distinctive improvements can be observed for C4.5-based ensembles for 7 of 9 cases, when compared to the models built using the original, unreduced data sets. VQNN-based ensembles have also been improved for 3 data sets: `cleveland`, `heart`, and `wine`. As the value of ε increases, the averaged sizes of the selected feature subsets become much smaller, but the improvement in terms of classification accuracy (over unreduced data) also becomes less evident. However, with such levels of reduction in data, this is expected.

It is important to point out that all the base classifiers employed in the experiments are trained using only the objects selected by the ε-bireducts, while the number of features is also greatly reduced in each case. Therefore, the accuracies of the learned models, when employed individually, may appear worse than those obtainable using the original, unreduced data. The experimental results demonstrate that, by combining the outputs of such individually weak base classifiers in an ensemble setting, the classification accuracy can be greatly improved.

The results also reveal a significant and pronounced effect upon the data set `libras`, which has 91 features and 15 possible class labels. Although the base classifiers (each trained on $< 7\%$ of features) can only achieve $\approx 40\%$ accuracy on average, their combined performance outperforms models learned using the unreduced data (with C4.5). This is a very positive indication that the proposed ensemble structure is effective, and that there is a good level of diversity within the discovered bireducts. The use of ensembles also helps avoid situations where minority classes may suffer from reduced representation in the final bireduct due to their high frequency of appearance in the clause lists. Furthermore, the value of ε not only has a direct impact on the number of covered objects $|Y|$,

TABLE V. ENSEMBLE CLASSIFICATION RESULTS FOR $\varepsilon = 0.1$ (90% INTENDED OBJECT COVERAGE)

Data Set	C4.5 Classification Accuracy (%)			VQNN Classification Accuracy (%)			Bireduct Coverage (%)	
	Ensemble	AVG Base	Unreduced	Ensemble	AVG Base	Unreduced	AVG($\frac{\|B\|}{\|\mathbb{A}\|}$)	AVG($\frac{\|Y\|}{\|\mathbb{U}\|}$)
glass	**68.25**	68.16	67.79	64.90	63.19	64.87	94.33	90.29
cleveland	52.25	52.43	53.85	**56.98**	53.85	52.22	45.57	89.90
heart	**81.48**	76.33	74.81	**78.89**	75.41	75.93	47.08	90.12
ionosphere	**89.13**	83.17	81.30	80.87	76.43	82.61	14.29	89.86
libras	**70.56**	48.42	64.72	63.61	49.44	65.28	6.91	90.12
sonar	**76.45**	65.43	74.57	74.95	67.07	76.00	9.33	89.85
water	**82.05**	78.18	81.28	78.97	78.33	81.79	14.49	90.03
wine	**94.31**	84.62	93.73	**93.86**	87.53	93.20	30.36	90.01
ecoli	79.73	79.91	80.61	84.52	84.75	84.48	63.75	90.58

TABLE VI. ENSEMBLE CLASSIFICATION RESULTS FOR $\varepsilon = 0.2$ (80% INTENDED OBJECT COVERAGE)

Data Set	C4.5 Classification Accuracy (%)			VQNN Classification Accuracy (%)			Bireduct Coverage (%)	
	Ensemble	AVG Base	Unreduced	Ensemble	AVG Base	Unreduced	AVG($\frac{\|B\|}{\|\mathbb{A}\|}$)	AVG($\frac{\|Y\|}{\|\mathbb{U}\|}$)
glass	**67.45**	65.32	65.43	63.18	62.76	68.23	69.44	80.27
cleveland	**55.57**	52.54	53.85	**56.23**	52.06	52.21	44.36	80.17
heart	74.82	73.37	74.82	75.93	70.37	75.93	32.54	80.25
ionosphere	**89.13**	79.65	81.30	79.13	72.83	83.04	11.57	80.19
libras	**71.67**	43.69	64.72	61.94	43.36	64.72	6.62	79.94
sonar	73.14	62.80	74.57	73.07	64.13	75.60	6.56	80.23
water	77.69	76.15	81.28	74.87	76.10	81.54	11.31	80.06
wine	88.17	78.53	93.73	85.29	78.68	93.20	27.36	80.02
ecoli	77.65	77.97	80.62	81.53	81.41	84.48	64.38	80.09

TABLE VII. ENSEMBLE CLASSIFICATION RESULTS FOR $\varepsilon = 0.3$ (70% INTENDED OBJECT COVERAGE)

Data Set	C4.5 Classification Accuracy (%)			VQNN Classification Accuracy (%)			Bireduct Coverage (%)	
	Ensemble	AVG Base	Unreduced	Ensemble	AVG Base	Unreduced	AVG($\frac{\|B\|}{\|\mathbb{A}\|}$)	AVG($\frac{\|Y\|}{\|\mathbb{U}\|}$)
glass	57.49	56.95	65.82	55.63	56.42	65.48	57.78	70.09
cleveland	**57.25**	54.17	50.59	**57.61**	54.15	54.94	37.50	70.07
heart	67.41	68.15	77.78	66.67	65.07	75.56	26.92	70.03
ionosphere	76.52	72.48	85.65	73.04	67.04	83.04	9.37	70.05
libras	62.50	41.50	70.83	58.61	40.58	67.78	5.80	70.06
sonar	**77.81**	66.33	73.12	75.45	66.14	76.93	6.80	70.09
water	76.38	73.33	83.09	78.17	76.00	81.54	11.84	70.12
wine	80.35	75.42	93.73	74.15	73.70	93.20	25.50	70.06
ecoli	72.26	71.54	80.62	74.06	73.47	84.48	50.75	70.45

but also on the number of selected features. This is important to note, as essentially any reduction achieved using a fuzzy-rough bireduct results in a sub-table of the original data. It is encouraging therefore to observe that an ensemble of such sub-tables has the ability to offer increased performance when compared to the use of unreduced data.

VI. CONCLUSION

This paper has presented a heuristic strategy for the purpose of identifying quality fuzzy-rough bireducts. The challenging task of simultaneous feature and instance selection or reduction has been tackled by employing the music-inspired harmony search algorithm. The notion of ε-bireduct is the key to identifying the desirable candidate solutions, as it helps as a guide to partially quantify the balance between the number of features and data instances to be formed in a given bireduct. The stochastic operations utilised by the proposed method help to generate multiple, similar quality bireducts, from which subsequent classifier ensembles can be constructed. The use of ε-bireduct-based classifier ensemble alleviates the loss of classification accuracy experienced by very compact bireducts, making the resultant system more accurate and robust.

Although promising, the present work offers room for improvement in a number of aspects. As it is a preliminary investigation, no attempt has been made to optimise the parameters and configurations for the employed methods (including the choice of $\mathcal{I}$, $\mathcal{T}$, and the fuzzy similarity relation). It can be expected that the performance of the proposed approach with optimisation would be even better than that presented here. Further, more in-depth, systematic comparative studies and rigorous statistical evaluation of the developed approaches (particularly with respect to the aspects regarding the discovery of fuzzy-rough bireducts from high dimensional data) remain active research.

It is worth noting that the complexity (in terms of both space and time) of the fuzzy discernibility function [14] has a high impact on the run-time efficiency of the bireduct search process. Alternative representations or computational procedures (i.e., the computation of the fuzzy-rough core [6]) may offer a way in which to identify more important fuzzy clauses. Also, it is interesting to investigate the relative classification accuracies associated with the selected bireducts as this would offer an insight into which sub-tables of features and data instances are more important for the generation of bireducts. Bireduct selection in a dynamic setting [11], [27] is also a very interesting topic for further exploration. Last but not least, it would be beneficial to investigate the diversity and stability of the bireduct-based classifier ensembles, as the

ensemble performance may be much improved if redundant or misleading members are identified and removed. Indeed, FS-inspired ensemble reduction methods [28] may provide a means for achieving such goals.

REFERENCES

[1] H. Liu and H. Motoda, *Computational Methods of Feature Selection (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 2007.

[2] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann, 2001, pp. 601–608.

[3] T. Kietzmann, S. Lange, and M. Riedmiller, "Incremental GRLVQ: Learning relevant features for 3D object recognition," *Neurocomputing*, vol. 71, no. 13-15, pp. 2868–2879, 2008.

[4] C. Shang and D. Barnes, "Fuzzy-rough feature selection aided support vector machines for mars image classification," *Computer Vision and Image Understanding*, vol. 117, no. 3, pp. 202–213, 2013.

[5] 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.

[6] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 4, pp. 824–838, 2009.

[7] R. Jensen and C. Cornelis, "Fuzzy-rough instance selection," in *IEEE International Conference on Fuzzy Systems*, 2010, pp. 1–7.

[8] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Norwell, MA, USA: Kluwer Academic Publishers, 1992.

[9] R. Jensen and Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*. Wiley-IEEE Press, 2008.

[10] D. Dubois and H. Prade, *Putting rough sets and fuzzy sets together*. Intelligent Decision Support, Kluwer Academic Publishers, Dordrecht,, 1992.

[11] S. Stawicki and D. Ślęzak, "Recent advances in decision bireducts: Complexity, heuristics and streams," in *Rough Sets and Knowledge Technology*, ser. Lecture Notes in Computer Science, P. Lingras, M. Wolski, C. Cornelis, S. Mitra, and P. Wasilewski, Eds. Springer Berlin Heidelberg, 2013, vol. 8171, pp. 200–212.

[12] N. Mac Parthalain and R. Jensen, "Simultaneous feature and instance selection using fuzzy-rough bireducts," in *IEEE International Conference on Fuzzy Systems*, 2013, pp. 1–8.

[13] R. Diao and Q. Shen, "Feature selection with harmony search," *IEEE Trans. Syst., Man, Cybern. B*, vol. 42, no. 6, pp. 1509–1523, 2012.

[14] R. Jensen, A. Tuson, and Q. Shen, "Finding rough and fuzzy-rough set reducts with SAT," *Information Sciences*, vol. 255, no. 0, pp. 100–120, 2014.

[15] J. Wróblewski, "Ensembles of classifiers based on approximate reducts," *Fundamenta Informaticae*, vol. 47, no. 3-4, pp. 351–360, Oct. 2001.

[16] S. H. Nguyen and H. S. Nguyen, "Pattern extraction from data," *Fundamenta Informaticae*, vol. 34, no. 1, pp. 129–144, 1998.

[17] Z. Zhou, *Ensemble Methods: Foundations and Algorithms*, ser. Chapman & Hall/CRC Data Mining and Knowledge Discovery Serie. Taylor & Francis, 2012.

[18] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search strategies for ensemble feature selection," *Information Fusion*, vol. 6, no. 1, pp. 83–98, 2005.

[19] D. Ślęzak and A. Janusz, "Ensembles of bireducts: Towards robust classification and simple representation," in *3rd International Conference on Future Generation Information Technology*, 2011, pp. 64–77.

[20] J. S. Olsson, "Combining feature selectors for text classification," in *Proceedings of the 15th ACM international conference on Information and knowledge management*, 2006, pp. 798–799.

[21] D. W. Opitz, "Feature selection for ensembles," in *Proceedings of 16th National Conference on Artificial Intelligence*, 1999, pp. 379–384.

[22] V. Torra and Y. Narukawa, *Modeling Decisions: Information Fusion and Aggregation Operators*. Springer, 2007.

[23] A. Frank and A. Asuncion, "UCI machine learning repository," 2010.

[24] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., ser. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Jun. 2005.

[25] R. Jensen and C. Cornelis, "Fuzzy-rough nearest neighbour classification," in *Transactions on Rough Sets XIII*, ser. Lecture Notes in Computer Science, J. Peters, A. Skowron, C.-C. Chan, J. Grzymala-Busse, and W. Ziarko, Eds. Springer Berlin Heidelberg, 2011, vol. 6499, pp. 56–72.

[26] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of K-fold cross-validation," *Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, Sep. 2004.

[27] R. Diao, N. Mac Parthaláin, and Q. Shen, "Dynamic feature selection with fuzzy-rough sets," in *IEEE International Conference on Fuzzy Systems*, Jun. 2013, pp. 1–7.

[28] R. Diao, F. Chao, T. Peng, N. Snooke, and Q. Shen, "Feature selection inspired classifier ensemble reduction," *IEEE Trans. Cybern.*, to appear.