Kernel-Based Fuzzy-Rough Nearest Neighbour Classification

Yanpeng Qu, Changjing Shang,
Qiang Shen, Neil Mac Parthaláin
Dept. of Computer Science
Aberystwyth University
Aberystwyth, Ceredigion, Wales. UK
Email: {yyq09, cns, qqs, ncm}@aber.ac.uk

Wei Wu
School of Mathematical Sciences
Dalian University ofTechnology
Dalian, 116024, China
Email: wuweiw@dlut.edu.cn

***Abstract*—Fuzzy-rough sets play an important role in dealing with imprecision and uncertainty for discrete and real-valued or noisy data. However, there are some problems associated with the approach from both theoretical and practical viewpoints. These problems have motivated the hybridisation of fuzzy-rough sets with kernel methods. Existing work which hybridises fuzzy-rough sets and kernel methods employs a constraint that enforces the transitivity of the fuzzy T-norm operation. In this paper, such a constraint is relaxed and a new kernel-based fuzzy-rough set approach is introduced. Based on this, novel kernel-based fuzzy-rough nearest-neighbour algorithms are proposed. The work is supported by experimental evaluation, which shows that the new kernel-based methods offer improvements over the existing fuzzy-rough nearest neighbour classifiers. The abstract goes here.**

***Keywords*—Fuzzy-rough sets; Fuzzy tolerance relation; Kernel theory; Nearest neighbour classification**

I. Introduction

Fuzzy-rough set theory [1] is a hybridisation of rough sets [2] and fuzzy sets [3], which is capable of dealing with imprecision and uncertainty in data. As a hybridisation of fuzzy set theory and rough sets, fuzzy-rough sets not only inherit the domain independence of rough sets, but also address the inability of rough sets in handling real-valued data. That is, fuzzy-rough sets provide a means to deal with discrete or real-valued noisy data (or a mixture of both) without the need for user-supplied thresholding or domain information. As such, this technique can be applied to regression as well as classification tasks. The only additional information required is in the form of fuzzy partitions for each feature which can be automatically derived from the data [4].

Kernel methods [5] have the ability to deal with non-linear models by mapping a given problem from the (low dimensional) input space onto a new (higher-dimensional) space via a non-linear transformation. The resulting structure of the classification task is then linearly separable. From another perspective, the kernel defines a similarity measure between two data objects and thus allows the utilisation of prior knowledge of the problem domain. More importantly, the kernel provides all of the information about the relative positions of the inputs in the feature space so that the associated learning algorithm is based only on the kernel function. Classification can be carried out without explicit use of the feature space.

The initial work on hybridising fuzzy-rough sets and kernel methods is presented in [6]. This work exploited the approach as described in [7], which explored the relationship between kernels and T-transitivity. In particular, kernel methods are integrated into fuzzy-rough sets (and more recently gaussian kernels [8]). In the work of [6], the concept of *kernelised fuzzy-rough sets* was proposed, in which kernel functions are employed to compute the fuzzy similarity relations between samples. One of the shortcomings of this approach however, is the fact that the fuzzy relations are limited only to T_{cos} equivalence relations [7] in an attempt to guarantee T-transitivity. However, as argued in [9], T-transitivity does not necessarily need to be enforced for fuzzy-rough sets [10], as the use of fuzzy tolerance relations [11] may be sufficient [12].

In this paper, an improved approach to using kernel methods with fuzzy-rough sets is presented. The reason why the proposed method is termed *kernel-based fuzzy-rough sets* rather than *kernelised fuzzy-rough sets*, is that in such a combination, kernels are employed as a special means to construct fuzzy tolerance relations. The framework of fuzzy-rough sets is preserved, whilst a statistical perspective is used in order to investigate the properties of kernels which may be suitable for integration into fuzzy-rough sets. To demonstrate the utility of the new kernel-based fuzzy-rough sets approach, a new form of nearest-neighbour classifier is proposed. This type of classifier also employs a vaguely-quantified rough set measure [13], which is robust in the presence of noisy data.

The remainder of this paper is structured as follows. The theoretical background is presented in Section 2 with a short review of existing methods. The proposed kernel-based fuzzy-rough set approach and the associated nearest neighbour algorithms are described in Section 3. The new kernel-based classifier is compared to others, with experimental results shown in Section 4. Finally, section 5 concludes the paper with a short discussion of future work.

II. Theoretical Background

A. Hybridisation of Rough Sets and Fuzzy Sets

The work on rough set theory (RST)[2] provides a methodology that can be employed to extract knowledge from a

domain in a concise way: It is able to minimise information loss whilst reducing the amount of knowledge involved. Central to rough set theory is the concept of indiscernibility. Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe) and $\mathbb{A}$ is a non-empty finite set of attributes so that $a : \mathbb{U} \longrightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that attribute a may take. For any $P \subseteq \mathbb{A}$, there exists an associated equivalence relation $IND(P)$:

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

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

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

where,

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

and,

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

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

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

$$\overline{P}X = \{x \mid [x]_P \cap X \neq \emptyset\}. \quad (6)$$

The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called a rough set.

The process described above although useful can only operate effectively on datasets containing discrete values. As most datasets contain real-valued attributes, a subjective judgement or threshold must therefore be employed in order for RST to operate on such data. The imposition of such a subjective threshold is however, contrary to the concept of domain independence of RST. An appropriate way of handling the problem of real-valued data is the use of fuzzy-rough sets (FRS) [1]. FRS offers a high degree of flexibility in enabling the vagueness and imprecision present in real-valued data to be modelled effectively.

Definitions for the fuzzy lower and upper approximations can be found in [10], [14], where a T-transitive fuzzy similarity relation is used to approximate a fuzzy concept X:

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x, y), \mu_X(y)) \quad (7)$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in \mathbb{U}} T(\mu_{R_P}(x, y), \mu_X(y)). \quad (8)$$

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

$$\mu_{R_P}(x, y) = T_{a \in P}\{\mu_{R_a}(x, y)\}. \quad (9)$$

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

$$\mu_{R_a}(x, y) = 1 - \frac{|a(x) - a(y)|}{|a_{max} - a_{min}|} \quad (10)$$

$$\mu_{R_a}(x, y) = \max\left(\min\left(\frac{(a(y) - (a(x) - \sigma_a))}{(a(x) - (a(x) - \sigma_a))}, \frac{((a(x) + \sigma_a) - a(y))}{((a(x) + \sigma_a) - a(x))}\right), 0\right) \quad (11)$$

where ${\sigma_a}^2$ is the variance of feature a. As these relations do not necessarily display T-transitivity, fuzzy transitive closure must be computed for each feature [9]. In other words, T-transitivity is not required for fuzzy-rough sets. Instead, fuzzy tolerance relations [11] can be employed to construct fuzzy-rough sets [12]. This technique is adopted in this paper also.

Note that formulas (7) and (8) are quite sensitive to noisy values, just like their crisp counterparts. Thus, the concept of vaguely-quantified rough set (VQRS) has been introduced in [13]. Following this approach, given a pair of fuzzy quantifiers (Q_u, Q_l), which are an increasing $[0, 1] \rightarrow [0, 1]$ mapping, the lower and upper approximation of X by R are defined by

$$\mu_{\underline{R_P}X}^{Q_u}(x) = Q_u\left(\frac{|R_P(x, y) \cap X|}{|R_P(x, y)|}\right) = Q_u\left(\frac{\sum_{y \in \mathbb{U}} \min(\mu_{R_P}(x, y), \mu_X(y))}{\sum_{y \in \mathbb{U}} \mu_{R_P}(x, y)}\right) \quad (12)$$

$$\mu_{\overline{R_P}X}^{Q_l}(x) = Q_l\left(\frac{|R_P(x, y) \cap X|}{|R_P(x, y)|}\right) = Q_l\left(\frac{\sum_{y \in \mathbb{U}} \min(\mu_{R_P}(x, y), \mu_X(y))}{\sum_{y \in \mathbb{U}} \mu_{R_P}(x, y)}\right), \quad (13)$$

The fuzzy set intersection is defined by the T-norm *min* and the fuzzy set cardinality by the sigma-count operation. As an important difference to (7) and (8), the VQRS approximations do not extend the classical rough set approximations, in a sense that when X and R are crisp, (12) and (13) may still be fuzzy.

B. Fuzzy-rough Nearest Neighbour Algorithm

A number of techniques have been developed for building fuzzy-rough nearest neighbour (FRNN) classifiers [15], [16]. Based upon such techniques, an approach which utilises the fuzzy upper and lower approximations to determine class membership is proposed in [17].

FRNN works by examining each of the decision classes in the training data in-turn. It computes the membership of a test object to the fuzzy lower and upper approximations of each class. These values are then compared with the highest existing values: $\mu_1(y)$ and $\mu_2(y)$. If the approximation membership

values for the currently considered class are higher, then both $\mu_1(y)$ and $\mu_2(y)$ are assigned these values and the class label is assigned to this test object. If not, the algorithm continues to iterate through all remaining decision classes. Classification accuracy is calculated by comparing the output with the actual class labels of the test objects.

An extension of FRNN is vaguely quantified rough nearest neighbour (FRNN-VQRS) [13] which employs (12) and (13), to determine class membership of test objects. The underlying learning mechanism is very similar to that of FRNN.

C. Classes of Kernels in Statistics

In a kernel algorithm, a mapping ϕ from the original space to a possibly high-dimensional space is employed to change the distribution of the data from nonlinear problem to linearly separable problem. By replacing the inner product with an appropriate kernel function, one can implicitly perform a nonlinear mapping to a high dimensional feature space without increasing the number of parameters. Consider the case of mapping an n-dimensional feature space to an m-dimensional feature space:

$$\phi : \mathbf{x} \to \phi(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n,\ \phi(\mathbf{x}) \in \mathbb{R}^m \tag{14}$$

A kernel denotes a function K such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y}). \tag{15}$$

In statistics, symmetric positive definite functions are called covariances. Hence, kernels are covariance-based in essence. From a statistics perspective, generally, two important classes of kernels are: stationary kernels and non-stationary kernels [18]. The work in this paper focuses on stationary kernels.

Stationary kernels $K(\mathbf{x}, \mathbf{y}) = K_S(\mathbf{x} - \mathbf{y})$ do not depend on the data object values themselves, but only on the lag vector separating the two objects **x** and **y**. Isotropic stationary kernels, which depend only on the norm of the lag vector, are most commonly used. For isotropic stationary kernels, the covariance form is:

$$K_{cov}(\mathbf{x}, \mathbf{y}) = K_I(\|\mathbf{x} - \mathbf{y}\|), \tag{16}$$

and the correlation form is

$$K_{cor}(\mathbf{x}, \mathbf{y}) = K_I(\|\mathbf{x} - \mathbf{y}\|)/K_I(0). \tag{17}$$

A *non-stationary* kernel $K(\mathbf{x}, \mathbf{y})$ is one which depends explicitly on the two data objects **x** and **y**. Note that a special kind of non-stationary kernel, called a *reducible kernel*, can be reduced to a stationary kernel.

III. Kernel-based Fuzzy-rough Nearest Neighbour Classification

A. Kernel-based Fuzzy-rough Sets

The relationship between T-transitivity and kernels has been explored [7]. It has been shown that any kernel $K : X \times X \to [0, 1]$, $K(x, x) = 1\ \forall x \in X$, is T_{cos}-transitive, where $T_{cos}(a, b) = max(ab - \sqrt{1 - a^2}\sqrt{1 - b^2}, 0)$. As an initial attempt, kernelised fuzzy-rough sets, which combine kernel methods with concepts from fuzzy-rough set theory, have been presented in [6], [8]. In this approach, kernels $K(x, y)$ are constrained such that they impose: a) reflexivity, b) symmetry, and c) T_{cos}-transitivity. Such kernels are employed to calculate the degree to which objects x and y are similar for every feature. The fuzzy lower and upper approximations in kernelised fuzzy-rough sets are defined by:

$$\mu_{\underline{K_P}X}(x) = \inf_{y \in \mathbb{U}} I_{cos}(K_P(x, y), \mu_X(y)) \tag{18}$$

$$\mu_{\overline{K_P}X}(x) = \sup_{y \in \mathbb{U}} T_{cos}(K_P(x, y), \mu_X(y)), \tag{19}$$

where, the implicator

$$I_{cos} = \begin{cases} 1, & a \le b \\ ab + \sqrt{(1 - a^2)(1 - b^2)}, & a > b \end{cases}.$$

However, as shown previously, in fuzzy-rough sets, T-transitivity is not necessarily displayed, and fuzzy tolerance relations may be sufficient [12]. Moreover, as (9), the fuzzy similarity relation induced by the subset of features P should be a combination by T-norm. Specifically, for kernelised fuzzy-rough sets, it is:

$$K_P(x, y) = T_{a \in P}\{\mu_{R_a}(x, y)\}. \tag{20}$$

In this case, the choice of a kernel function becomes limited. This is due to the fact that not many kernel functions can be denoted by a T-norm-based combination of reflexive functions. For instance, the Gaussian kernel employed in [8] and [6] is workable, because for: $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$

$$\exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\theta}\right) = \prod_{i=1}^{n} \exp\left(-\frac{(x_i - y_i)^2}{\theta}\right) \tag{21}$$

and because its product is still a T-norm. However, for most kernels, such as the rational quadratic kernel and the wave kernel (see below), this property may not hold. In order to address these problems, kernel-based fuzzy-rough sets are proposed in this paper.

In geometry, the inner product of two vectors is the projection of one onto another. Actually, the square of the norm distance in Hilbert space can be expressed by the inner product. In this case, the inner product can measure the similarity between the images of two features by mapping into a Hilbert space. Therefore, given a non-empty set $\mathbb{U}$ and a kernel function K being reflexive, (that is $K(x, x) = 1$), for an arbitrary fuzzy concept X, the lower and upper approximations of a kernel-based fuzzy-rough set can be defined as:

$$\mu_{\underline{R_P^K}X}(x) = \inf_{y \in \mathbb{U}} I(\mu_{K_P}(x, y), \mu_X(y)) \tag{22}$$

$$\mu_{\overline{R_P^K}X}(x) = \sup_{y \in \mathbb{U}} T(\mu_{K_P}(x, y), \mu_X(y)). \tag{23}$$

It is important to note that the framework of fuzzy-rough sets remains intact using the definition described in this paper. In other words, the kernel methods play a special role in calculating the fuzzy tolerance relations. It is because of this

fact that the term *kernel-based fuzzy-rough sets* (KFRS) is employed here rather than *kernelised fuzzy-rough sets*.

As well as fuzzy-rough sets, the corresponding the lower and upper approximations of the kernel-based vaguely quantified rough set (KVQRS) can be also be defined:

$$\mu_{\underline{R_P^K}X}^{Q_u}(x) = Q_u\left(\frac{|R_P^K(x,y)\cap X|}{|R_P^K(x,y)|}\right) = Q_u\left(\frac{\sum\limits_{y\in\mathbb{U}}\min(\mu_{K_P}(x,y),\mu_X(y))}{\sum\limits_{y\in\mathbb{U}}\mu_{K_P}(x,y)}\right) \quad (24)$$

$$\mu_{\overline{R_P^K}X}^{Q_l}(x) = Q_l\left(\frac{|R_P^K(x,y)\cap X|}{|R_P^K(x,y)|}\right) = Q_l\left(\frac{\sum\limits_{y\in\mathbb{U}}\min(\mu_{K_P}(x,y),\mu_X(y))}{\sum\limits_{y\in\mathbb{U}}\mu_{K_P}(x,y)}\right), \quad (25)$$

where, $\mu_{K_P}(x,y)$ is induced by the subset of features P and kernel function K:

$$\mu_{K_P}(x,y) = T_{a\in P}\{\phi(a(x))\cdot\phi(a(y))\} = T_{a\in P}\{K(a(x),a(y))\} = T_{a\in P}\{K_a(x,y)\}. \quad (26)$$

As established previously, all isotropic stationary kernels in correlation form (17) are suitable for being integrated into KFRS. A collection of certain commonly used isotropic stationary kernels in correlation form are listed as follows:

- Gaussian kernel: $K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\theta}\right)$
- Exponential kernel: $K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|}{\theta}\right)$
- Rational quadratic kernel: $K(\mathbf{x},\mathbf{y}) = 1 - \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\|\mathbf{x}-\mathbf{y}\|^2+\theta}$
- Wave kernel: $K(\mathbf{x},\mathbf{y}) = \frac{\theta}{\|\mathbf{x}-\mathbf{y}\|}\sin(\frac{\|\mathbf{x}-\mathbf{y}\|}{\theta})$.

It is worth noting that for specific non-stationary kernels, the reflexivity holds also. For instance, the non-stationary kernel [18],

$$K(\mathbf{x},\mathbf{y}) = \frac{\|\mathbf{x}\| + \|\mathbf{y}\| - \|\mathbf{x}-\mathbf{y}\|}{2\sqrt{\|\mathbf{x}\|\|\mathbf{y}\|}}, \quad (27)$$

is reflexive. This kernel is also reducible.

B. Kernel-based Fuzzy-rough Nearest Neighbour Classification

The present work initially aims to investigate the combination of kernel methods with conventional fuzzy-rough nearest neighbour approaches (FRNN and FRNN-VQRS) [17]. The resulting combined learning algorithm is outlined in Figure 1. As with FRNN, the rationale behind this algorithm is that the the lower and the upper approximations of each decision class (calculated by means of the nearest neighbours of a test object y) will provide helpful clues to predict the membership of a test object to any given class. The complexity of this algorithm is: $O(|\mathcal{C}|\cdot 2|\mathbb{U}|)$.

KFRNN($\mathbb{U}$,$\mathcal{C}$,y)
$\mathbb{U}$, the training set; $\mathcal{C}$, the set of decision classes;
y, the object to be classified.

(1) $N \leftarrow$ get Nearest Neighbour(y,k)
(2) $\mu_1(y)\leftarrow 0, \mu_2(y)\leftarrow 0, Class\leftarrow\emptyset$
(3) $\forall X\in\mathcal{C}$
(4) **if**$(\mu_{\underline{R_P^K}X}(y)\geq\mu_1(y)\&\&\mu_{\overline{R_P^K}X}(y)\geq\mu_2(y))$
(5) $\overline{Class}\leftarrow X$
(6) $\mu_1(y)\leftarrow\mu_{\underline{R_P^K}X}(y), \mu_2(y)\leftarrow\mu_{\overline{R_P^K}X}(y)$
(7) **output** $Class$

Fig. 1. The kernel-based fuzzy-rough nearest neighbour algorithm

The algorithm in Figure 1 can be further adapted to perform kernel-based vaguely quantified rough nearest neighbour (KFRNN-VQRS) classification, by replacing: $\mu_{\underline{R_P^K}X}(y)$ and $\mu_{\overline{R_P^K}X}(y)$ with $\mu_{\underline{R_P^K}X}^{Q_u}(y)$ and $\mu_{\overline{R_P^K}X}^{Q_l}(y)$, respectively.

In statistics, the stationary property is often mathematically assumed to describe the ability of ensuring that a random process maintains the same probabilistic characteristics such as mean, variance and autocorrelation. Typically, non-stationary and bifurcated regimes are always observed in the case of datasets with class imbalance [19]. Class imbalance occurs when one or more classes are over or under represented as a total number of objects of the whole dataset. Thus, for datasets which suffer from class imbalance, the distribution of and the similarity between the objects may be non-stationary. From this point of view, if a dataset is extremely imbalanced, the use of non-stationary kernels would be more appropriate. Further investigation would help to confirm this - see conclusion section for further discussion.

IV. Experimental Results

This section presents an experimental evaluation of the proposed algorithms (KFRNN-FRS and KFRNN-VQRS) using two different kernels: 1) an isotropic stationary kernel, and 2) a non-stationary kernel. Nine benchmark datasets obtained from[20] are employed. These datasets are small-to-medium in size, containing between 178 and 683 objects with feature numbers ranging from 6 to 279.

TABLE I
Evaluation Datasets

Dataset	Objects	Attributes
Arrhythmia	452	279
Glass	214	9
Heart	270	13
Liver	345	6
Sonar	208	60
Water 2	390	38
Water 3	390	38
Wine	178	14
Wisconsin	683	9

For the evaluation described here, k is set at an initial value of 10 for FRNN-VQRS and KFRNN-VQRS. For FRNN

the relation given in equation (10) is used. For the kernel-based methods, an exponential kernel (used as the isotropic stationary kernel), and the non-stationary kernel of (27) are employed. In the FRNN and KFRNN approaches, the Kleene-Dienes *T*-norm is used to implement the implicator, which is defined by $I(x, y) = \max(1 - x, y)$. The FRNN-VQRS and KFRNN-VQRS approaches are implemented with $Q_l = Q_{(0.1,0.6)}$ and $Q_u = Q_{(0.2,1.0)}$, according to the general formula

$$Q_{(\alpha,\beta)}(x) = \begin{cases} 0, & x \leq \alpha \\ \frac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \alpha \leq x \leq \frac{\alpha+\beta}{2} \\ 1 - \frac{2(x-\beta)^2}{(\beta-\alpha)^2}, & \frac{\alpha+\beta}{2} \leq x \leq \beta \\ 1, & \beta \leq x \end{cases} . \quad (28)$$

Stratified 10 $\times$ 10-fold cross-validation (10-FCV) is employed for result validation. In 10-FCV, the original dataset is partitioned into 10 subsets. Of these 10 subsets, a single subset is retained as the testing data for the model, and the remaining 9 subsets are used for training. The cross-validation process is then repeated 10 times (the number of folds). The 10 sets of results are then aggregated via averaging to produce a single model estimation. The advantage of 10-FCV over random sub-sampling is that all objects are used for both training and testing, and each object is used for testing only once per fold. The stratification of the data prior to its division into folds ensures that each class label (as far as possible) has equal representation in all folds, thus helping to alleviate bias/variance problems. In order to investigate the level of 'fit' of these models, the root mean squared error (RMSE) measure is used. The RMSE is the squared root of the variance of the residuals. It indicates the absolute fit of a model to the data and how close the observed data objects are to the model predicted values. Note that, RMSE is an absolute measure. As the squared root of a variance, RMSE can be viewed as the standard deviation of the unexplained variance. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and is a generally accepted criterion for assessing fit, if the purpose of the resulting model is for prediction. In addition, conventional classification accuracy is also used to assess the performance of learnt classifiers.

A. Performance Evaluation

A comparison of the kernel-based and FRFS-based nearest-neighbour techniques is shown in Tables II and III, where sKFRNN and nonsKFRNN stand for stationary kernel-based fuzzy-rough nearest neighbour and non-stationary kernel-based fuzzy-rough nearest neighbour methods, respectively. Correspondingly, stationary kernel-based vaguely quantified rough nearest neighbour and non-stationary kernel-based vaguely quantified rough nearest neighbour approaches are denoted by sKFRNN-VQRS and nonsKFRNN-VQRS, respectively.

Note that the *Water2* and *Water3* datasets, both suffer from class imbalance. In particular, the ratios of data between different classes are: 312 objects to 78 objects for classes 1 and 2 for *water2*, and 378 objects for class 1 and 12 objects for class 2 for *water3*. It can be seen from the experimental results that the non-stationary kernel based approaches, nonsKFRNN and nonsKFRNN-VQRS, consistently achieved the highest classification accuracy and smallest RMSE over all other methods for these datasets. However, for the rest datasets which are not highly imbalanced, the differences between the performance of two new approaches and existing techniques are not that significant. In fact, for *Liver* and *Glass*, both of sKFRNN and sKFRNN-VQRS reached the highest classification accuracy and smallest RMSE values. For *Heart*, *Sonar*, *Wine* and *Wisconsin*, the competitiveness of the kernel-based fuzzy-rough nearest neighbour classifiers are obvious. Putting these results together, it is clear that the present work helps to improve the quality of fuzzy-rough nearest neighbour classifiers.

B. Statistical Analysis

To further evaluate the kernel-based fuzzy-rough techniques, a paired t-test with significance level of 0.05 has been carried out. The baseline references for the tests are the results obtainable from FRNN and FRNN-VQRS classification, respectively. This is done in order to ensure that the aforementioned results are not discovered by chance. The statistical significance results are shown in Table IV and V, where the symbols 'v', '*' and '-' indicate that the results are statistically better, worse, or have no statical significance.

The results once again demonstrate that the kernel-based approaches achieve best performances overall for *Water2*, *Water3*, *Heart* and *Arrythmia*, etc. In particular, compared to FRNN, the non-stationary kernel-based methods are shown to be statistically better than the other methods for these datasets. Note however that no statistical differences are found amongst all tested algorithms for the *Glass*, *Liver*, *Wine* and *Wisconsin* datasets. Only a single dataset (*sonar*), do the proposed techniques occasionally return a result which is statistically worse than that attainable using FRNN. This may well be due to the use of the fuzzy quantifiers in FRNN-VQRS, though further experimental evaluation is required in order to verify this. For FRNN-VQRS, the kernel-based modification leads to a statistically similar performance generally. sKFRNN and nonsKFRNN are better than FRNN-VQRS statistically for the *Glass* and *Sonar* datasets, although worse for the *Arrythmia* and *Heart* datasets. This may possibly be caused by the existence of noisy data in *Arrythmia* and *Heart*.

V. Conclusion

This paper has presented a new technique for the hybridisation of fuzzy-rough sets and kernel methods, called kernel-based fuzzy-rough sets (KFRS). In contrast to previous work, the T_{cos}-transitivity constraint is relaxed. The only remaining constraint which the proposed approach imposes is reflexivity.

Whilst attempting to identify suitable kernels, the properties are analysed from statistics perspective. It has been demonstrated that all isotropic stationary kernels in the correlation form of (17) are suitable for use with the KFRS

TABLE II
COMPARISON BETWEEN FRNN, SKFRNN, NONSKFRNN

Dataset	FRNN		sKFRNN		nonsKFRNN	
	Accy.	RMSE	Accy.	RMSE	Accy.	RMSE
Arrhythmia	54.67	0.21	52.33	0.22	55.43	0.21
Glass	73.54	0.29	76.24	0.28	73.21	0.29
Heart	76.63	0.43	76.59	0.42	73.00	0.43
Liver	62.81	0.50	63.70	0.49	61.06	0.49
Sonar	85.25	0.43	84.69	0.43	85.95	0.43
Water 2	75.41	0.46	73.95	0.46	83.00	0.39
Water 3	67.87	0.40	65.79	0.39	75.77	0.34
Wine	97.47	0.20	98.15	0.21	96.12	0.21
Wisconsin	96.38	0.19	96.65	0.20	96.65	0.20

TABLE III
COMPARISON BETWEEN FRNN-VQRS, SKFRNN-VQRS, NONSKFRNN-VQRS

Dataset	FRNN-VQRS		sKFRNN-VQRS		nonsKFRNN-VQRS	
	Accy.	RMSE	Accy.	RMSE	Accy.	RMSE
Arrhythmia	60.40	0.21	59.42	0.21	62.13	0.20
Glass	68.95	0.27	74.08	0.26	66.57	0.29
Heart	82.19	0.35	82.41	0.35	75.81	0.43
Liver	66.26	0.48	67.19	0.48	67.72	0.48
Sonar	79.38	0.37	80.83	0.36	75.98	0.41
Water 2	79.59	0.39	80.49	0.40	84.92	0.33
Water 3	73.18	0.37	72.67	0.38	80.26	0.31
Wine	97.14	0.10	96.97	0.12	94.22	0.15
Wisconsin	96.69	0.16	96.16	0.17	96.81	0.15

TABLE IV
STATISTICAL SIGNIFICANCE USING PAIRED T-TEST FOR FRNN

Dataset	FRNN	sKFRNN	nonsKFRNN	sKFRNN-VQRS	nonsKFRNN-VQRS
Arrythmia	-	-	-	v	v
Glass	-	-	-	-	-
Heart	-	-	-	v	-
Liver	-	-	-	-	-
Sonar	-	-	-	-	*
Water2	-	-	v	v	v
Water3	-	-	v	-	v
Wine	-	-	-	-	-
Wisconsin	-	-	-	-	-

TABLE V
STATISTICAL SIGNIFICANCE USING PAIRED T-TEST FOR FRNN-VQRS

Dataset	FRNN-VQRS	sKFRNN	nonsKFRNN	sKFRNN-VQRS	nonsKFRNN-VQRS
Arrythmia	-	*	*	-	-
Glass	-	v	-	-	-
Heart	-	*	-	-	-
Liver	-	-	-	-	-
Sonar	-	-	v	-	-
Water2	-	-	-	-	-
Water3	-	-	-	-	-
Wine	-	-	-	-	-
Wisconsin	-	-	-	-	-

approach. Two kernel-based fuzzy-rough set classifiers: kernel-based fuzzy-rough nearest neighbour (KFRNN) and kernel-based vaguely quantified rough nearest neighbour (KFRNN-VQRS) have been introduced. The experimental results over 9 datasets, show that the new methods are effective, and that they generally outperform the original techniques.

Topics for further investigation include the impact of the choice of kernel, connectives and quantifiers on performance. Also, (and as mentioned previously), the relationship between the class imbalance of datasets and the statistical property of kernels is a worthwhile avenue of exploration. Considering hierarchical classification, another further extension to this

work would be to examine how KFRS performs for the task of feature selection [4].

REFERENCES

[1] D. Dubois and H. Prade, "Putting rough sets and fuzzy sets together," *Intelligent Decision Support*, pp. 203–232, 1992.

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

[3] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

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

[5] J. Shawe-Taylor and N. Christianini, *Kernel method for pattern analysis*. Cambridge: Cambridg University Press, 2004.

[6] Q. Hu, D. Chen, D. Yu, and W. Pedrycz, "Kernelised fuzzy rough sets," in *4th International Conference, Rough Sets and Knowledge Technology*, 2009, pp. 304–311.

[7] B. Moser, "On the T-transitivity of kernels," *Fuzzy Sets and Systems*, vol. 157, pp. 1787–1796, 2006.

[8] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu, "Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications," *International Journal of Approximate Reasoning*, vol. 51, no. 4, pp. 453–471, 2010.

[9] M. D. Cock, C. Cornelis, and E. Kerre, "Fuzzy rough sets: The forgotten step," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 1, pp. 121–130, 2007.

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

[11] M. Das, M. Chakraborty, and T. Ghoshal, "Fuzzy tolerance relation, fuzzy tolerance space and basis," *Fuzzy Sets and Systems*, vol. 97, pp. 361–369, 1998.

[12] R. Jensen and C. Cornelis, "Fuzzy-rough instance selection," in *Proceedings of the 19th International Conference on Fuzzy Systems*, 2010, pp. 1776–1782.

[13] C. Cornelis, M. D. Cock, and A. Radzikowska, "Vaguely quantified rough sets," *Lecture Notes in Artificial Intelligence*, vol. 4482, pp. 87–94, 2007.

[14] 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.

[15] H. Bian and L. Mazlack, "Fuzzy-rough nearest-neighbor classification approach," in *Proceeding of the 22nd International Conference of the North American Fuzzy Information Processing Society*, 2003, pp. 500–505.

[16] M. Sarkar, "Fuzzy-rough nearest neighbors algorithm," *Fuzzy Sets and Systems*, vol. 158, pp. 2123–2152, 2007.

[17] R. Jensen and C. Cornelis, "A new approach to fuzzy-rough nearest neighbour classification," in *Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing*, 2008, pp. 310–319.

[18] M. Genton, "Classes of kernels for machine learning: a statistics perspective," *Journal of Machine Learning Research*, vol. 2, pp. 299–312, 2001.

[19] C. Drioli and F. Avanzini, "Non-modal voice synthesis by low-dimensional physical models," in *Proceeding of the 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2003.

[20] C. Blake and C. Merz, "UCI repository of machine learning databases," 1998, university of California, Irvine, School of Information and Computer Sciences.