Fuzzy-rough approaches for mammographic risk analysis

MacParthaláin, Neil Seosamh; Jensen, Richard; Shen, Qiang; Zwiggelaar, Reyer

Published in:
Intelligent Data Analysis

DOI:
10.3233/IDA-2010-0418

Publication date:
2010

Citation for published version (APA):
MacParthaláin, N. S., Jensen, R., Shen, Q., & Zwiggelaar, R. (2010). Fuzzy-rough approaches for mammographic risk analysis. *Intelligent Data Analysis*, *14*(2), 225-244. https://doi.org/10.3233/IDA-2010-0418

Fuzzy-rough approaches for mammographic risk analysis

Neil Mac Parthaláin*, Richard Jensen, Qiang Shen, and Reyer Zwiggelaar
{ncm, rkj, qqs, rrz}@aber.ac.uk
Department of Computer Science,
Aberystwyth University,
Ceredigion, Wales, UK, SY23 3DB

Abstract

The accuracy of methods for the assessment of mammographic risk analysis is heavily related to breast tissue characteristics. Previous work has demonstrated considerable success in developing an automatic breast tissue classification methodology which overcomes this difficulty. This paper proposes a unified approach for the application of a number of rough and fuzzy-rough set methods to the analysis of mammographic data. Indeed this is the first time that fuzzy-rough approaches have been applied to this particular problem domain. In the unified approach detailed here feature selection methods are employed for dimensionality reduction developed using rough sets and fuzzy-rough sets. A number of classifiers are then used to examine the data reduced by the feature selection approaches and assess the positive impact of these methods on classification accuracy. Additionally, this paper also proposes a new fuzzy-rough classifier based on the nearest neighbour classification algorithm. The novel use

of such an approach demonstrates its efficiency in improving classification accuracy for mammographic data, as well as considerably removing redundant, irrelevant, and noisy features. This is supported with experimental application to two well-known datasets. The overall result of employing the proposed unified approach is that feature selection can identify only those features which require extraction. This can have the positive effect of increasing the risk assessment accuracy rate whilst additionally reducing the time required for expert scrutiny, which in-turn means the risk analysis process is potentially quicker and involves less screening.

Keywords: Fuzzy-rough sets, Rough sets, Feature selection, Classification, Mammographic risk assessment.

1 Introduction

Breast cancer is a major health issue, and the most common amongst women in the EU. It is estimated that 8–13% of all women will develop breast cancer at some point during their lives [6], [16]. Furthermore, in the EU and US, breast cancer is recognised as the leading cause of death of women in their 40s [6], [7], [16]. Although increased incidence of breast cancer has been recorded, so too has the level of early detection through the screening of potential occurence using mammographic imaging and expert opinion. However, even expert radiologists can sometimes fail to detect a significant proportion of mammographic abnormalities. In addition, a large number of detected abnormalities are usually discovered to be benign following medical investigation.

Existing mammographic Computer Aided Diagnosis (CAD) systems [21, 32] concentrate on the detection and classification of mammographic abnormalities. As breast tissue density increases however, the effectiveness of such systems in detecting mammographic abnormalities is reduced significantly. Also, it is known that there is a strong correlation between mammographic breast tissue density and the risk of development of breast cancer. Automatic classification which has the ability to consider tissue density when searching for mammographic abnormalities is therefore highly desirable.

This paper presents two novel ideas. The first is a unified approach which employs a number of rough and fuzzy-rough approaches to deal with mammographic data. In particular, this method considers each step from feature extraction through to data classification, although this paper focuses primarily on the latter two steps. The second idea is a new nearest-neighbour classifier based on fuzzy-rough sets which is one of a number of fuzzy-rough classifier components that can be 'plugged-into' the previously described unified approach. The use of the unified approach results in a reduction in the number of misleading or redundant image features and the new classifier demonstrates improved classification accuracy.

The remainder of the paper is structured as follows. An overview of related work is presented in section 2, this forms the basis for the work demonstrated later in the paper. A unified fuzzy-rough framework is proposed in section 3 along with an examination of the methods employed in the dimensionality reduction, and classification phases. In section 4, the new fuzzy-rough nearest

neighbour (FRNN) classification algorithm is introduced, and a worked example is presented. Section 5 demonstrates the application of a number of classifiers including the fuzzy-rough nearest neighbours (FRNN) approach, to two mammographic datasets. Additionally, the experimental setup is discussed in this section, and comparative results are presented for a number of dimensionality reduction and classifier approaches within the framework presented earlier. Section 6 concludes the paper with a short discussion of future work.

2 Background

It is important to emphasise that the problem under consideration in this paper is *mammographic risk analysis* rather than mammographic diagnosis from images, an area where many publications have been written with respect to the application of machine learning techniques [1], [5], [18], [35]. Mammographic risk analysis, involves the extraction of mammographic breast density information from images which is then used to assess how likely a woman is to develop breast cancer. The basic steps involved are outlined in Fig.1, with detailed background described in [30]. The initial stages involve the segmentation and filtering of the mammographic images: all mammograms are pre-processed to identify the breast region and remove image background, labels, and pectoral muscle areas. This segmentation step results in a very minor loss of skin-line pixels in the breast area, however these pixels are not required for tissue estimation.

Then, a feature extraction step is performed, where the fuzzy c-means (FCM)

algorithm [4] is employed which results in the division of the breast into two clusters. A co-occurrence matrix (which is essentially a 2D histogram) is then used to derive a feature set which results in 10 features to describe morphological characteristics and 216 for the texture information (226 total). This feature set is then labelled using the consensus opinion of 3 experts to manually assign a label to each actual mammographic image using the BIRADS [2] classification. This consensus is determined where the classification for a given mammogram, which two or three radiologists agreed upon (majority vote) is selected as the 'consensus class'. If all experts classified a single mammogram differently, the median value is chosen as consensus opinion. The divergence in the opinion of the experts, is a major factor which often frustrates the use of automatic methods. This highlights the need to remove inter-observer (inter-operator) variability through the development of more autonomous approaches.

The approach outlined in Fig.1 is used as a starting point for the unified approach which is proposed in this paper where the existing classification step is replaced with a dimensionality reduction phase and a classification phase. The existing extracted feature set is used, as is the consensus expert labelling of the data.

3 Unified Fuzzy-Rough Approach

A unified framework such as that shown in Fig.2 is adopted to simplify the way in which knowledge can be efficiently learned from the (mammographic) training data, and therefore applied to real-world risk assessment problems. In this work,

the focus lies in the implementation of rough and fuzzy-rough techniques for the dimensionality reduction and classifier learner steps. The approach for the feature extraction step employed in this paper is documented in [30], however there is no reason why future work could not include a fuzzy-rough method to accomplish this in an effort to unify the underlying mathematical approach (see conclusion for further discussion).

Efficient, and in particular accurate classification of mammographic imaging is of high importance. Any improvement in accuracy for automatic mammographic classification systems can result in a reduction in the amount of required expert analysis thus improving the time taken to perform breast tissue risk assessment. The data in mammographic imaging is real-valued and as mentioned previously can be noisy. Clearly, any classifier employed must therefore have the ability to deal with such data. Discrete methods require that the real-valued data is discretised and thus result in information loss, however the methods described in this paper require no discretisation and use only the information contained within the data. The following sections describe the fuzzy-rough dimensionality and classification methods which are used within the proposed unified framework.

3.1 Dimensionality Reduction

In this work, feature selection (FS) is utilised as the dimensionality reduction technique. This allows the identification of a minimal feature subset from a problem domain while retaining both a suitably high accuracy and the semantics entailed by the original features. In many real world problems, FS is necessary

due to the level of noisy, irrelevant or misleading features. By removing these factors, techniques for learning from data can benefit greatly. A detailed review of FS techniques devised for classification tasks can be found in [12], [24], [27], and [28]. The focus of this paper however lies in the implementation of FS techniques which are based exclusively on rough sets and fuzzy rough sets.

The work on rough set theory (RST) [31], offers a formal methodology that can be employed to reduce the dimensionality of datasets, as a preprocessing step to assist any chosen modeling method for learning from data. It assists in identifying and selecting the most information-rich features in a dataset. This is achieved without transforming the data, whilst simultaneously attempting to minimise information loss during the selection process. In terms of computational effort, this approach is highly efficient, as it is based on simple set operations. This makes it suitable as a preprocessor for techniques that are much more complex. In contrast to statistical correlation-reduction approaches [13], RST requires no human input or domain knowledge other than the given datasets. Perhaps most importantly though, it retains the underlying semantics of the data, which results in data models that are more transparent to human scrutiny.

At the heart of the rough set approach is the concept of indiscernibility. Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe) and $\mathbb{A}$ is a non-empty finite set of attributes so that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that a can take. For any $P \subseteq \mathbb{A}$, there exists an associated equivalence relation $IND(P)$:

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

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

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

where,

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

and,

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

where A and B are families of sets.

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:

$$\underline{P}X = \{x \mid [x]_P \subseteq X\} \tag{5}$$

$$\overline{P}X = \{x \mid [x]_P \cap X \neq \emptyset\} \tag{6}$$

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

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

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

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

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

$$k = \gamma_P(Q) = \frac{\mid POS_P(Q) \mid}{\mid \mathbb{U} \mid} \tag{10}$$

The reduction of attributes or selection of survival features can be realised through the comparison of equivalence relations generated by sets of attributes. Attributes are removed such that the reduced set provides identical predictive capability of the decision feature or features as that of the original or unreduced set of features.

The QuickReduct algorithm [10] shown in Fig. 3 searches for a minimal subset without exhaustively generating all possible subsets. The search begins with an empty subset, attributes which result in the greatest increase in the rough set dependency value are added iteratively. This process continues until

the search produces its maximum possible dependency value for that dataset ($\gamma_{\mathbb{C}}(\mathbb{D})$). Note that this type of hill-climbing search does not guarantee a minimal subset and may only discover a local minimum.

3.2 Rough and fuzzy-rough feature selection

Unfortunately, one of the main disadvantages of the classical rough set methodology is its inability to deal with real-valued data unless the data is discretised which can result in information loss. One particular extension which has been proposed to address this shortcoming is the tolerance rough set model (TRSM) [38]. other extensions such as variable precision rough sets (VPRS) [45] deal with misclassification of objects rather than real-valued data.

This paper utilises two approaches: one which is rough set-based and another which is based on fuzzy-rough sets, both of which have the ability to deal with real-valued data. The first approach implements a version of tolerance rough sets [38] which also takes advantage of the information in the boundary region or region of uncertainty [29]. The second utilises fuzzy-rough sets [14] which extend the rough set approach outlined previously.

Unlike the classical rough set methodology TRSM employs a similarity relation to minimise data as opposed to the indiscernibility relation, thus allowing a degree of 'fuzziness'. This allows a relaxation in the way equivalence classes are considered. This flexibility allows a blurring of the boundaries of the former rough or crisp equivalence classes and objects may now belong to more than one tolerance class. Suitable similarity relations must be defined for each feature, although a common definition can be used for all features if applicable. A

standard measure for this purpose, given in [38], is:

$$SIM_a(x, y) = 1 - \frac{\mid a(x) - a(y) \mid}{\mid a_{max} - a_{min} \mid} \tag{11}$$

where a is a considered feature, and a_{max} and a_{min} denote the maximum and minimum values of a respectively.

When considering the case where there is more than one feature, the defined similarities must be combined to provide an overall measure of similarity of objects. For a subset of features, P, this can be achieved in many ways including the following approaches:

$$(x, y) \in SIM_{P,\tau} \iff \prod_{a \in P} SIM_a(x, y) \geq \tau \tag{12}$$

$$(x, y) \in SIM_{P,\tau} \iff \frac{\sum_{a \in P} SIM_a(x, y)}{\mid P \mid} \geq \tau \tag{13}$$

where τ is a global similarity threshold and determines the required level of similarity for inclusion within a tolerance class. This framework allows for the specific case of traditional rough sets by defining a suitable similarity measure (e.g. complete equality of features and the use of equation (12)) and threshold ($\tau = 1$). An algorithm can be formulated which uses this framework to search for subsets in the same way that the classical rough set QuickReduct does. Further detail of this approach can be found in [29].

The requirement of rough set theory to rely on discrete data implies an objectivity in the data that is simply not present. For example, consider an

attribute *Blood Pressure* in a medical dataset. In the real world, this is a real-valued measurement but for the purposes of RST must be discretized into a small set of labels such as *Normal*, *High*, etc. Subjective judgments are therefore required to establish boundaries for objective measurements.

A more appropriate way of handling this problem is the use of *fuzzy-rough* sets [14]. Subjective judgments are not entirely removed as fuzzy set membership functions still need to be defined. However, the method offers a high degree of flexibility when dealing with real-valued data, enabling the vagueness and imprecision present to be modelled effectively.

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. Vagueness arises due to a lack of distinction or hard boundaries in the data itself. This is typical of human communication and reasoning. Rough sets can be said to model ambiguity resulting from a lack of information through set approximations.

Definitions for the fuzzy lower and upper approximations can be found in [34], where a T-transitive fuzzy similarity relation is used to approximate a fuzzy concept X:

$$\mu_{\underline{R_P}X}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x,y), \mu_X(y)) \tag{14}$$

$$\mu_{\overline{R_P}X}(x) = \sup_{y \in \mathbb{U}} T(\mu_{R_P}(x,y), \mu_X(y)) \tag{15}$$

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

$$\mu_{R_P}(x,y) = T_{a \in P}\{\mu_{R_a}(x,y)\} \tag{16}$$

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

$$\mu_{R_a}(x, y) = 1 - \frac{|a(x) - a(y)|}{|a_{max} - a_{min}|} \tag{17}$$

$$\mu_{R_a}(x, y) = \exp(-\frac{(a(x) - a(y))^2}{2{\sigma_a}^2}) \tag{18}$$

$$\mu_{R_a}(x, y) = \max(\min(\frac{(a(y) - (a(x) - \sigma_a))}{(a(x) - (a(x) - \sigma_a))}, \frac{((a(x) + \sigma_a) - a(y))}{((a(x) + \sigma_a) - a(x))}, 0) \tag{19}$$

where ${\sigma_a}^2$ is the variance of feature a. As these relations do not necessarily display T-transitivity, the fuzzy transitive closure can be computed for each attribute.

In a similar way to the original crisp rough set approach, the fuzzy positive region [23] can be defined as:

$$\mu_{POS_{R_P}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{R_P} X}(x) \tag{20}$$

An important issue in data analysis is the discovery of dependencies between attributes. This is of particular significance for feature selection and pattern classification. The fuzzy-rough dependency degree of $\mathbb{D}$ on the attribute subset P can be defined as:

$$\gamma'_P(\mathbb{D}) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_{R_P}(\mathbb{D})}(x)}{|\mathbb{U}|} \tag{21}$$

A fuzzy-rough reduct R is defined as a subset of features which preserves the dependency degree of the entire dataset, i.e. $\gamma'_R(\mathbb{D}) = \gamma'_{\mathbb{C}}(\mathbb{D})$. Based on this, a fuzzy-rough QuickReduct algorithm can be constructed that uses equation

(21) as shown in to gauge subset quality when searching for a minimal reduct. It is this algorithm which is used in this paper to generate reducts for mammographic data.

3.3 Classifier Learning

The work presented here addresses the classification of mammographic imaging. This requires the learning of the classifier employed in the overall system (see Fig.2). A number of existing classifiers as well as a new hybrid fuzzy-rough classifier are examined. These include: FNN [26], a fuzzy version of the well-known kNN algorithm [15]; FRNN-O a fuzzy-rough ownership function based classifier [36, 41]; and VQNN a nearest neighbour (NN) classifier based on the vaguely quantified rough set model [11].

In the previous approach [30] shown in Fig.1, conventional crisp classifier learners were employed for the classification of the mammographic data – kNN, C4.5 [33], and a combined Bayesian estimation approach type classifier [15]. In this paper a number of hybrid fuzzy set and rough set-based classifiers have been employed to classify the mammographic data. Each classifier algorithm is discussed in further detail in this section.

The kNN algorithm assigns a test object to the decision class most common among its 'k nearest neighbours', i.e., the k training objects that are closest to the test object. An extension of the kNN algorithm to fuzzy set theory (FNN) was introduced in [26]. It allows partial membership of an object to different classes, and also takes into account the relative importance (proximity) of each neighbour with respect to the test instance. However, as correctly argued in [36],

the FNN algorithm has problems dealing adequately with insufficient knowledge. In particular, when every training pattern is far removed from the test object, and hence there are no suitable neighbours, the algorithm is still forced to make clear-cut predictions. This is because the sum of the predicted membership degrees to the various decision classes is always required to be equal to 1.

For the purposes of FNN, the extent to which an unclassified object y belongs to class X is defined as:

$$\mu_X(y) = \sum_{x \in N} \mu_R(x, y)\mu_X(x) \tag{22}$$

where N is the set of object y's k-nearest neighbours and $\mu_R(x, y)$ is the fuzzy similarity of y and object x. In the traditional fuzzy kNN approach, this is defined in the following way:

$$\mu_R(x, y) = \frac{||y - x||^{-2/(m-1)}}{\sum_{j \in N} ||y - j||^{-2/(m-1)}} \tag{23}$$

where $||\cdot||$ denotes the Euclidean norm, and m is a parameter that controls the overall weighting of this fuzzy similarity. The FNN algorithm employs these definitions to determine the extent to which an object y belongs to each class, typically classifying y to the class with the highest resulting membership. The complexity of this algorithm for the classification of one test pattern is $O(|\mathbb{U}| + k \cdot |C|)$,

Initial attempts to combine the FNN algorithm with concepts from fuzzy rough set theory were presented in [36, 41] (here denoted FRNN-O). In these papers, a fuzzy-rough ownership function is constructed that attempts to handle both "fuzzy uncertainty" (caused by overlapping classes) and "rough uncer-

tainty" (caused by insufficient knowledge, i.e. attributes, about the objects). All training objects influence the ownership function, and hence no decision is required as to the number of neighbours to consider, although there are other parameters that must be defined for its successful operation.

Note that the approach does not use fuzzy lower or upper approximations to determine class membership unlike the method proposed in this paper. The fuzzy-rough ownership function was defined as:

$$\tau_X(y) = \frac{\sum_{x \in \mathbb{U}} \mu_R(x,y)\mu_X(x)}{|\mathbb{U}|} \tag{24}$$

This can be modified to consider only the k nearest neighbours as follows:

$$\tau_X(y) = \frac{\sum_{x \in N} \mu_R(x,y)\mu_X(x)}{|N|} \tag{25}$$

where N is the set of object y's k-nearest neighbours. When $k = |\mathbb{U}|$ the original definition is obtained as illustrated in eqn. (25). The fuzzy similarity is determined by

$$\mu_R(x,y) = \exp(-\sum_{a \in \mathbb{C}} \kappa_a (a(y) - a(x))^{2/(m-1)}) \tag{26}$$

where m controls the weighting of the similarity (as in FNN) and κ_a is a parameter that decides the bandwidth of the membership, defined as

$$\kappa_a = \frac{|\mathbb{U}|}{2\sum_{x \in \mathbb{U}} ||a(y) - a(x)||^{2/(m-1)}} \tag{27}$$

For FRNN-O, initially a parameter κ_a is calculated for each attribute and all memberships of decision classes for test object y are set to zero. Next, the

weighted distance of y from all objects in the universe is computed and used to update the class memberships of y via eqn. (24). Finally, when all training objects have been considered, the algorithm outputs the class with the highest membership. The complexity of the algorithm is $O(|\mathbb{C}||\mathbb{U}| + |\mathbb{U}| \cdot (|\mathbb{C}| + |C|))$. This method still requires a choice of parameter m, which plays a similar role to that in FNN.

Equations (14) and (15) have been conceived with the purpose of conserving the traditional lower and upper approximations in mind. Indeed, when X and R_P are both crisp, it can be verified that the original crisp rough set definitions are recovered. Note in particular how the inf and sup operations play the same role as the $\forall$ and $\exists$ quantifiers of the classical rough sets approach, and how a change in a single element can thus have a large impact on (14) and (15). This makes fuzzy-rough sets equally susceptible to noisy data (which is difficult to rule out in real-life applications) as their crisp counterparts.

To make up for this shortcoming, the work in [11] proposed to soften the universal and existential quantifier by means of vague quantifiers like *most* and *some*. Mathematically, the vague quantifiers were modeled in terms of Zadeh's notion of a regularly increasing fuzzy quantifier Q: an increasing $[0, 1] \to [0, 1]$ mapping that satisfies the boundary conditions $Q(0) = 0$ and $Q(1) = 1$.

Examples of fuzzy quantifiers can be generated by means of the following

parametrised formula, for $0 \leq \alpha < \beta \leq 1$, and x in $[0,1]$,

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

For instance, $Q_{(0.1,0.6)}$ and $Q_{(0.2,1)}$ might be used respectively to reflect the vague quantifiers *some* and *most* from natural language.

Once a couple (Q_l, Q_u) of fuzzy quantifiers is fixed, the Q_l-upper and Q_u-lower approximation of a fuzzy set A under a fuzzy relation R are defined by

$$\mu_{\underline{R_P}X}^{Q_u}(y) = Q_u\left(\frac{|R_P y \cap X|}{|R_P y|}\right) \tag{29}$$

$$\mu_{\overline{R_P}X}^{Q_l}(y) = Q_l\left(\frac{|R_P y \cap X|}{|R_P y|}\right) \tag{30}$$

for all y in $\mathbb{U}$. In other words, an element y belongs to the lower approximation of X if most of the elements related to y are included in X. Likewise, an element belongs to the upper approximation of X if some of the elements related to y are included in X. Notice that when X and R_P are a crisp set and a crisp equivalence relation respectively, the approximations may still be non-crisp.

The algorithm given in Fig. 4 can be adapted to perform VQRS-based nearest neighbours (VQNN) classification by replacing $\mu_{\underline{R_P}X}(y)$ and $\mu_{\overline{R_P}X}(y)$ with $\mu_{\underline{R_P}X}^{Q_u}(y)$ and $\mu_{\overline{R_P}X}^{Q_l}(y)$. The computational complexity of this approach is similar to that of classical rough set approach.

4 Fuzzy-Rough Nearest Neighbours

This section concentrates on the description of the novel fuzzy-rough nearest neighbour algorithm. The need for such a new classification technique arises

from the fact that although the FRNN-O algorithm proposed in [36] uses a fuzzy-rough framework, no use is made of the fuzzy upper and lower approximations to determine class membership. This has prompted the development of an approach which was built upon the existing fuzzy-rough techniques which had been applied successfully to the feature selection problem [23]. As both the FS problem and the classification problem are similar is many ways, the motivation was therefore quite clear.

The intuitive basis for the approach is that the lower and the upper approximation of a decision class, calculated by means of the nearest neighbours of a test object y, provide good clues to predict the membership of the test object to that class. Thus, by calculating the upper and lower approximation of a given decision class these can be employed as a metric for the test object in determining class membership.

4.1 FRNN Algorithm

The membership of a test object y to each (crisp or fuzzy) decision class is determined via the calculation of the fuzzy lower and upper approximation. The algorithm outputs the decision class with the resulting best fuzzy lower and upper approximation memberships. More specifically, if the membership of y to the fuzzy lower approximation of class C is high, it means that all of $y's$ neighbours belong to class C, while a high membership value of the fuzzy upper approximation of C indicates that at least one neighbour or neighbours belong to that class. The algorithm iterates through all of class concepts (X) in the training data. The decision class which results in the highest upper and lower

approximation membership values is assigned to the test object.The complexity of the algorithm is $O(|C| \cdot (2|\mathbb{U}|))$.

Although the parameter k (number of nearest neighbours to consider) is not required, it can be incorporated into the algorithm by replacing line (2) with "$N \leftarrow$ getNearestNeighbours (y, k)". As $\mu_{R_P}(x, y)$ becomes smaller, x tends to have only a minor influence on $\mu_{\underline{R_P}X}(y)$ and $\mu_{\overline{R_P}X}(y)$.

The algorithm works by examining each of the classes in the training data in-turn. It computes the membership of a test object to the fuzzy upper and lower approximations. These values are then compared with the highest existing values $\mu_1(y)$ and $\mu_1(y)$. If the approximation membership values for the currently considered class are higher, then both $\mu_1(y)$ and $\mu_1(y)$ are assigned these values and the class label is assigned to this test object. If not, the algorithm continue to iterate through all remaining decision classes. Classification accuracy is calculated by comparing the output with the actual class labels of the test objects.

4.2 Worked example

In order to demonstrate the application of the algorithm, a small worked example is presented. This example employs a dataset with 3 real-valued conditional attributes (a,b, and c) and a single crisp discrete-valued decision attribute (q) as the *training* data, shown in Table 1. A further dataset shown in Table 2 containing 2 objects is used as the *test* data to be classified, again with the same number of conditional and decision attributes.

Referring to the FRNN algorithm described in the previous section, the first

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

Table 1: Example training data

Object	a	b	c	q
t1	0.3	-0.3	0	no
t2	-0.3	-0.4	-0.3	yes

Table 2: Example test data

step is to calculate the fuzzy upper and lower approximations for all decision classes. In Table 1 there are 4 objects and as noted previously a decision attribute which has 2 classes ($\{yes\}$, and $\{no\}$).

Using the fuzzy similarity measure as defined in (17) the similarity of each test object is compared to all of the objects in the training data. For instance, consider the training object $t1$:

$$\mu_{R\{P\}}(t1,1) = T(\mu_{R\{a\}}(t1,1), \mu_{R\{b\}}(t1,1), \mu_{R\{c\}}(t1,1)) = 0$$
$$\mu_{R\{P\}}(t1,2) = T(\mu_{R\{a\}}(t1,2), \mu_{R\{b\}}(t1,2), \mu_{R\{c\}}(t1,2)) = 0.16$$
$$\mu_{R\{P\}}(t1,3) = T(\mu_{R\{a\}}(t1,3), \mu_{R\{b\}}(t1,3), \mu_{R\{c\}}(t1,3)) = 0.83$$
$$\mu_{R\{P\}}(t1,4) = T(\mu_{R\{a\}}(t1,4), \mu_{R\{b\}}(t1,4), \mu_{R\{c\}}(t1,4)) = 0.40$$

These similarity values can then be used to generate the lower and upper approximations. Note that the fuzzy connectives chosen for this example are the Lukasiewicz t-norm ($\max(x + y - 1, 0)$), and Lukasiewicz fuzzy implicator

$(\min(1 - x + y), 1)$.

For the decision concept $X = yes$ these are:

$$\mu_{\underline{R_P}}\, X(t1) = \inf_{y \in \mathbb{U}} \{I(\mu_{R_P}(t1, y), \mu_X(y)\}$$

$$= \inf\{I(0,1), I(0.16,0), I(0.83,0), I(0.4,1)\} = 0.14$$

and,

$$\mu_{\overline{R_P}}\, X(t1) = \sup_{y \in \mathbb{U}} \{I(\mu_{R_P}(t1, y), \mu_X(y)\}$$

$$= \sup\{T(0,1), T(0.16,0), T(0.83,0), T(0.4,1)\} = 0.84$$

Similarly for the decision concept $X = no$:

$$\mu_{\underline{R_P}}\, X(t1) = \inf\{I(0,0), I(0.16,1), I(0.83,1), I(0.4,0)\} = 0.16$$

$$\mu_{\overline{R_P}}\, X(t1) = \sup\{T(0,0), T(0.16,1), T(0.83,1), T(0.4,0)\} = 0.86$$

With reference once again to the FRNN algorithm in Fig.4, it can be seen that the upper and lower approximation membership values for test object $t1$ for the class label $X = no$ are higher than those for when $X = yes$. The algorithm will therefore classify $t1$ as belonging to the class $X = no$. The procedure can be

repeated for training object $t2$ which results in upper and lower approximation values for $X = no$:

$$\mu_{\underline{R_P}}\, X(t2) = \inf\{I(0.6,1), I(0.6,0), I(0.17,0), I(0.17,1)\} = 0.4$$

$$\mu_{\overline{R_P}}\, X(t2) = \sup\{T(0.6,1), T(0.6,0), T(0.17,0), T(0.17,1)\} = 0.6$$

And, $X = yes$:

$$\mu_{\underline{R_P}}\, X(t2) = \inf\{I(0.6,0), I(0.6,1), I(0.17,1), I(0.17,0)\} = 0.4$$

$$\mu_{\overline{R_P}}\, X(t2) = \sup\{T(0.6,0), T(0.6,1), T(0.17,1), T(0.17,0)\} = 0.6$$

In this case, both upper and lower approximation membership values for each of the classes $X = no$ and $X = yes$ are identical. However because of line 6 of the FRNN algorithm, $t2$ will be classified as belonging to $X = yes$.

5 Experimentation

In this section the results of applying the previously described classifiers and FS preprocessors are presented. Initially the classifiers are applied to the unreduced extracted feature data - i.e. data on which FS has not been performed, see Fig.5. Classification is then performed on data which has been reduced by two previously described FS preprocessors DMTRS [29], and FRFS [23].

The results are then assessed for both FS methods used in conjunction with each of the individual classifiers. Additionally the results obtained in this paper are briefly compared with those reported in [30].

5.1 Experimental Setup

The are two datasets considered in this paper, and both are available in the public domain: the Mammographic Image Analysis Society (MIAS) database [39], and the Digital Database of Screening Mammography (DDSM) [19]. The MIAS dataset is composed of Medio-Lateral-Oblique (MLO) left and right mammograms from 161 women (322 objects). Each mammogram object is represented by 281 features extracted using the process detailed in [30]. The spatial resolution of the images is $50\mu m \times 50\mu m$ and quantized to 8 bits with a linear optical density in the range $0 - 3.2$.

The DDSM database provides four mammograms, comprising left and right Medio-Lateral-Oblique (MLO) and left and right Cranio-Caudal (CC) views, for most women. To avoid bias only the right MLO mammogram for each woman is selected. The dataset contains 832 mammograms (objects) and again 281 features obtained in the same manner as those for the MIAS dataset above.

The class labels for each mammogram are assigned by three experts consensus opinion as described previously in section 2. There are four discrete labels ranging from 1 to 4 relating to the BIRADS classification [2], where 1 represents a breast that is entirely fatty and 4 represents a breast that is extremely dense.

For the FRFS preprocessor the fuzzy similarity employed is defined in eqn.(19) along with the Łukasiewicz t-norm ($\max(x+y-1,0)$) and the Łukasiewicz fuzzy

implicator ($\min(1 - x + y, 1)$). It has been shown that these work particularly well when used for fuzzy-rough feature selection [23].

The DM-TRS preprocessor used 4 different tolerance values (τ) – 0.97 and 0.98 for the MIAS dataset, while for the DDSM dataset the values 0.98 and 0.99 were chosen. These were the values that empirically demonstrated the best level of dimensionality reduction for each of the datasets respectively.

For each of the classifier learners the value of k is initialised as 30 and then decremented by 1 each time, resulting in 30 experiments for each dataset. Such a wide range of values for k ensures a comprehensive exploration and comparison of each of the classifiers. Cross validation of 10×10-fold cross-validation (10-fold CV) is performed for each experiment. Note that the k parameter is essential only for FNN and FRNN-O and is not required for the other classifier learners. However, for ease of comparison, the other approaches have been adapted such that a k value can be specified. This is achieved by calculating the test objects k nearest neighbours rather than using all of the objects in the training set. For FNN and FRNN-O, m is set to 2. The VQNN approach was implemented using the commonly adopted $Q_l = Q_{(0.1,0.6)}$ and $Q_u = Q_{(0.2,1.0)}$, according to the general formula in equation (28).

For the new classifier approach, although there are no parameters to tune, decisions about which fuzzy relations and implicators must still be made. For the purpose of the experimentation documented in this paper, the fuzzy relation given in eqn. (17) was chosen for simplicity. In the FRNN approach, the min t-norm and the Kleene-Dienes implicator I (defined by $I(x, y) = \max(1 - x, y)$)

were used.

Initially, the three classification techniques described previously as well as the new FRNN technique are applied to the *unreduced* datasets and some results are then generated. The dimensionality of the data is then reduced and a summary of the average classification values achieved for each FS method is used to compare the methods.

5.2 Unreduced data

The classification accuracy results for the unreduced data are presented in this section. This was achieved by applying each of the four classifiers to both of the datasets which gives a background against which to make subsequent comparative studies.

Considering the classification accuracy results illustrated in Fig. 6, it can be seen that there is little variation in the performance for the MIAS dataset. The FRNN-O approach seems to have a slight advantage, however this is only in the order of 2-6% for all values of k. The results for the DDSM dataset tell a slightly different story with VQNN achieving a small but clear advantage. FNN also appears to marginally outperform FRNN, and FRNN-O methods follow a similar trend to that of VQNN. Generally, as the number of objects in the dataset increases, so too does the potential for measurement noise. The noise-tolerant characteristics of VQNN and the fact that the DDSM dataset has many more objects than the MIAS dataset may explain why VQNN performs particularly well in this case.

It is important to note at this point that the levels of performance shown

for the FRNN approach are of little importance in this section as the data prior to reduction with FS contains much redundancy, irrelevance, and noise.

5.3 Reduced Data

In this section the results of classifying the MIAS and DDSM datasets following feature selection are presented. Classification accuracy results are provided for both DMTRS and FRFS, again using both 10-fold CV. In Table 3, the subset sizes obtained following FS are presented. It is interesting to note that a substantial level of dimensionality reduction is achieved for both approaches. A reduction of 97.15% and 97.5% were achieved for the MIAS dataset, while the DDSM dataset (Table 4) achieved 97.15%, and 98.22%.

Orig No. of feats	DMTRS (τ=0.97)	DMTRS (τ=0.98)	FRFS
281	8	7	7

Table 3: Subset sizes - MIAS dataset after FS

Orig No. of feats.	DMTRS (τ=0.98)	DMTRS (τ=0.99)	FRFS
281	8	5	8

Table 4: Subset sizes - DDSM dataset after FS

The results presented here illustrate the classification accuracies obtained when using DMTRS as a FS preprocessing step. There are a total of four diagrams (Fig. 7 and 8), two of which represent the tolerance values for the MIAS dataset (0.97 and 0.98), and the remaining two represent the values for the DDSM dataset (0.98 and 0.99).

The results shown in Fig. 9 are those obtained when applying the classifiers to the data following the application of FRFS to reduce the data.

Perhaps the most obvious aspect of the results demonstrated here is the increase in classification accuracy for all classifiers following the use of FS. The advantages of applying FS are manifold, however in this case the level of dimensionality reduction and the aforementioned increase in classification accuracy are borne out in Figs. 7–9.

5.4 Comparative Investigation I: classifying unreduced data

As clearly demonstrated in Figs. 7–9 employing either method for FS results in a significant increase in classification accuracy. Importantly, the newly proposed FRNN technique performs best for both the MIAS and DDSM datasets, with the VQNN approach closely mirroring the performance of FNN. FRNN-O also seems to show similar accuracy for some values of k to FRNN but fails to do so consistently.

Figs. 7 – 9 present the classification accuracy results following the application of both the FRFS and DMTRS feature selection pre-processors. What is most noticeable about these results is the overall increase in classification accuracy when FS has been employed. This not only highlights the level of redundant features in the original (unreduced) dataset, but also the ability of fuzzy-rough FS methods to reduce the data dimensionality considerably.

For ease of comparison, the classification results of figs.7–9 have been summarised in Tables 5 and 6. Note that this summary is of *a*verage classification accuracy values. It is interesting that the subset sizes obtained for each FS ap-

proach. For example in Table II, the DMTRS approach achieves a subset sizes of 7 and 8 for the MIAS dataset. In Table 5, it can be seen that there is little difference in average classification accuracy between each of the tolerance values for DMTRS. Similarly, FRFS produces average classification results which are comparable with those of DMTRS for all classifiers. For the DDSM dataset however the DMTRS method manages better classification accuracies than FRFS for τ=0.98 for all classifiers except FRNN-O. Indeed the standard deviation values for this DMTRS subset is also lower than that achieved by FRFS. For the subset selected when τ=0.99, which is of size 5 compared to that of FRFS which is 8, there is little to separate FRFS and DMTRS in terms of average classification accuracy despite the greater level of dimensionality reduction.

Classifier	FRFS	st.dev	DMTRS ($\tau = 0.97$)	st.dev	DMTRS ($\tau = 0.98$)	st.dev
FRNN	86.99	6.85	86.69	7.16	86.30	7.07
FNN	75.78	8.65	71.18	10.13	71.61	10.03
FRNN-O	82.21	7.42	75.77	8.77	75.78	8.53
VQNN	76.85	8.34	80.85	8.10	80.75	7.98

Table 5: MIAS - Average classification accuracy, and standard deviation results

Classifier	FRFS	st.dev	DMTRS ($\tau = 0.98$)	st.dev	DMTRS ($\tau = 0.99$)	st.dev
FRNN	82.60	7.98	84.85	6.75	84.52	7.36
FNN	72.98	9.43	74.07	8.83	72.08	10.29
FRNN-O	81.14	8.69	77.39	7.67	74.14	9.44
VQNN	72.81	9.13	77.05	8.08	75.20	9.13

Table 6: DDSM - Average classification accuracy, and standard deviation

5.5 Comparative Investigation II: comparison with current state-of-the-art

When comparing the results obtained for this paper with those of [30], which represents the current state-of-the-art in automated mammographic breast density classification, it is clear that there is a significant improvement in classification accuracy. In [30], for the MIAS dataset classification rates of 77%, 72%, and 86% are achieved respectively for each of the classifier learners employed - namely SFS+kNN, C4.5, and a Bayesian classifier (although this is an approach which combines the previous two methods). Leave-one-out cross validation (LOOCV) is employed for cross validation in the paper in question, and k=7, for the kNN classifier. Additionally, the SFS+kNN approach employs a 'wrapper' type FS approach to select a subset of size 9 for MIAS and 9 also for the DDSM data.

Both DMTRS and FRFS feature selection approaches achieve results of 8 and 7 for MIAS and 5 and 8 for DDSM. Both of these approaches find smaller subset sizes when compared to the approach noted above whilst simultaneously leading to a significant increase in both average classification accuracy values albeit using 10-fold CV - see Fig. 7 - 9 . As demonstrated previously, more optimum values can be achieved for individual values of k (Figs. 7 - 9) rather than considering only those average classification accuracy results. Preliminary work which employs LOOCV also demonstrates that the unified approach adopted in this paper achieves results similar to those for 10-fold CV documented here, however such a comprehensive investigation is beyond the scope of this paper.

Both FS techniques employed in this paper are data-driven and do not re-

quire any normalisation or transformation of the data. In the work of [30] however, the data has to be normalised prior to the application of wrapper FS using kNN. This may have the effect of information loss since it involves subjective human intervention when dealing with the data.

Considering the FRNN results obtained in this paper for the MIAS dataset, classification accuracies of 91.4%, 90.28%, and 90.81%, were achieved for DMTRS($\tau = 0.97$), DMTRS($\tau = 0.98$), and FRFS reduced data respectively. Indeed, if the results from Table 5 are examined, it can be seen that even the average classification accuracies are considerably better in most cases than those obtained in [30].

For the DDSM dataset where classification accuracies of 70%, 72%, and 77% have been achieved in the previous work [30], considerably improved results have also been obtained using the new fuzzy and fuzzy-rough methods - 89.24%, 88.51%, and 85.84%. Again the average classification results of Table 6 reflect what has also been demonstrated in the case of the MIAS dataset.

Once again, although it is acknowledged that the classifiers were learned using 10-fold CV, the performance increases of the unified approach are emphasised by the subset size results and the increase in classification accuracy following the application of FS.

6 Conclusion

This paper has demonstrated the application of fuzzy-rough methods to data for mammographic risk analysis. It has also introduced a new NN classifica-

tion approach and demonstrated how this can be applied for the analysis of mammographic data. In particular, it has demonstrated how the classification accuracy for mammographic risk-analysis can be increased significantly by employing fuzzy classifiers which have the ability to handle real-valued data.

Most importantly however, the value of adopting a unified approach has been highlighted. This is clearly shown in the large improvement of classification accuracy over the unreduced data for all classifier methods and also the significant reduction in dimensionality, which has a direct impact on the time taken to classify mammographic density. The use of FS to identify information-rich features whilst minimising feature measurement noise from the many initially extracted features is important as it can then be used as an indicator to identify the same information in previously unseen mammograms thus, reducing the time needed in extracting many irrelevant, redundant and noisy features. Increases in classification accuracy for diagnosis means a benefit not only for the patient but also a reduction in expert analysis thus the minimising inter-observer variation. Additionally, correct initial identification of breast density can potentially mean that further additional screening of the same woman is not required, reducing the physical demands and stresses of further examination.

Areas for future work include the application of an unsupervised FS approach to the unlabeled MIAS and DDSM data and compare the classification results with those of the FS approaches used here. Also, a closer examination of the feature extraction process, especially in-terms of how rough and fuzzy-rough approaches could be applied to the extraction of features from mammographic

images is an area which requires further exploration. A further more in-depth evaluation, of the various fuzzy implicators, membership functions, and similarity measures which can employed for both the FS and classification phases of the approach is also required. The experimental evaluation in this paper utilised those which had been previously available and further improvements in performance may be realised through the use of different measures. Application of the unified FS/classifier approach could also be extended to other problem domains, such as forensic evidence [25], industrial systems monitoring [22], or heterogeneous data found in e.g. microarray analysis [3].

Acknowledgements

The authors wish to thank both the referees and the editor for their invaluable advice in revising this paper.

References

[1] F. Aghdasi, R.K. Ward, J. Morgan-Parkes, and B. Palcic. Feature selection for classification of mammographic microcalcification clusters. Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 58–59, 1993.

[2] American College of Radiology. Illustrated Breast Imaging Reporting and Data System BIRADS. 3rd Edition, American College of Radiology, 1998.

[3] D. Berrar, B. Sturgeon, I. Bradbury, C.S. Downes, and W. Dubitzky. Integration of microarray data for a comparative study of classifiers and iden-

tification of marker genes. Proceedings of the 4th International Conference on Critical Assessment of Microarray Data Analysis (CAMDA03), Durham, North Carolina, USA, pp. 147-162, 2004.

[4] J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[5] M. Brady, and R. Highnam. Mammographic Image Analysis, Kluwer Series on Medical Image Understanding, 1999.

[6] F. Bray, P. McCarron, and D.M. Parkin. The changing global patterns of female breast cancer incidence and mortality. Breast Cancer Research, vol. 6, pp. 229–239, 2004.

[7] S. Buseman, J. Mouchawar, N. Calonge, and T. Byers. Mammography screening matters for young women with breast carcinoma. Cancer, vol. 97, no. 2, pp. 352–358, 2003.

[8] M.A. Carreira-Perpinñán. Continuous latent variable models for dimensionality reduction and sequential data reconstruction. PhD thesis, University of Sheffield, UK. 2001.

[9] T-H. Cheng, C-P. Wei, and V.S. Tseng. Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. 19th IEEE International Symposium on Computer-Based Medical Systems, pp.165–170, 2006.

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

[11] C. Cornelis, M. De Cock, A.M. Radzikowska. Vaguely Quantified Rough Sets. Lecture Notes in Artificial Intelligence 4482, pp. 87–94, 2007.

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

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

[14] D. Dubois and H. Prade. Putting rough sets and fuzzy sets together. Intelligent Decision Support, Kluwer Academic Publishers, pp. 203–232. 1992.

[15] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification, 2nd ed. John Wiley and Sons, New York, 2001.

[16] Eurostat. Health statistics atlas on mortality in the European Union. Official Journal of the European Union, 2002.

[17] B.S. Everitt. An Introduction to Latent Variable Models, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1984.

[18] A. Hassanien. Fuzzy rough sets hybrid scheme for breast cancer detection, Image and Vision Computing, vol. 25, no. 2, pp. 172–183 2007.

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

[20] Q. Hu, Z. Xie, and D. Yu. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition, vol. 40, pp. 3509–3521, 2007.

[21] iCAD Second Look "http://www.icadmed.com" *accessed: 10/06/2008*

[22] Q. Shen and R. Jensen: Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring. Pattern Recognition 37(7): 1351-1363 (2004).

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

[24] R. Jensen and Q. Shen. Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches. IEEE PressWiley & Sons. 2008

[25] R. Jensen and Q. Shen. Feature Selection for Aiding Glass Forensic Evidence Analysis. To appear in Intelligent Data Analysis, vol. 13, no.5, 2009.

[26] J.M. Keller, M.R. Gray and J.A. Givens. A fuzzy K-nearest neighbor algorithm. IEEE Trans. Systems Man Cybernet., vol. 15, no. 4, pp. 580-585, 1985.

[27] H. Liu, and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining Springer International Series in Engineering and Computer Science, vol. 454, 1998.

[28] H. Liu and H. Motoda (eds), Computational Methods of Feature Selection, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2008.

[29] N. Mac Parthaláin, and Q. Shen. Exploring the boundary region of tolerance rough sets for feature selection. Pattern Recognition vol. 42 , no. 5 , pp. 655–667, 2009.

[30] 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.

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

[32] R2 ImageChecker, "http://www.r2tech.com" *accessed: 10/06/2008*

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

[34] 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.

[35] M. Roffilli. Advanced machine learning techniques for digital mammography, Technical Report UBLCS-2006-12, University of Bologna (Italy), 2006.

[36] M. Sarkar. Fuzzy-Rough nearest neighbors algorithm. Fuzzy Sets and Systems, vol. 158, pp. 2123–2152, 2007.

[37] Q. Shen and A. Chouchoulas. A rough-fuzzy approach for generating classification rules. Pattern Recognition, vol. 35, no. 11, pp. 2425-2438, 2002.

[38] A. Skowron, and J. Stepaniuk, Tolerance Approximation Spaces, Fundamenta Informaticae, vol. 27, pp. 245–253, 1996.

[39] 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 211–221, 1994.

[40] E.C.C. Tsang, D. Chen, D.S. Yeung, X-Z Wang, J. Lee. Attributes Reduction Using Fuzzy Rough Sets, IEEE Transactions on Fuzzy systems, vol. 16, no. 5, pp. 1130–1141, 2008.

[41] X. Wang, J. Yang, X. Teng and N. Peng. Fuzzy-Rough Set Based Nearest Neighbor Clustering Classification Algorithm. Lecture Notes in Computer Science, vol. 3613/2005, pp. 370–373, 2005.

[42] Y. Yang and J.O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the 14th International Conference on Machine Learning (ICML 97), pp. 412–420, 1997.

[43] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, vol. 13, no. 1, pp. 44–49, 1998.

[44] S. Zhao, and E.C.C. Tsang. On fuzzy approximation operators in attribute reduction with fuzzy rough sets, Information Sciences, vol. 178, pp.3163–3176, 2008.

[45] W. Ziarko, Variable Precision Rough Set Model, Journal of Computer and System Sciences, vol. 46, no. 1, pp. 39–59, 1993.

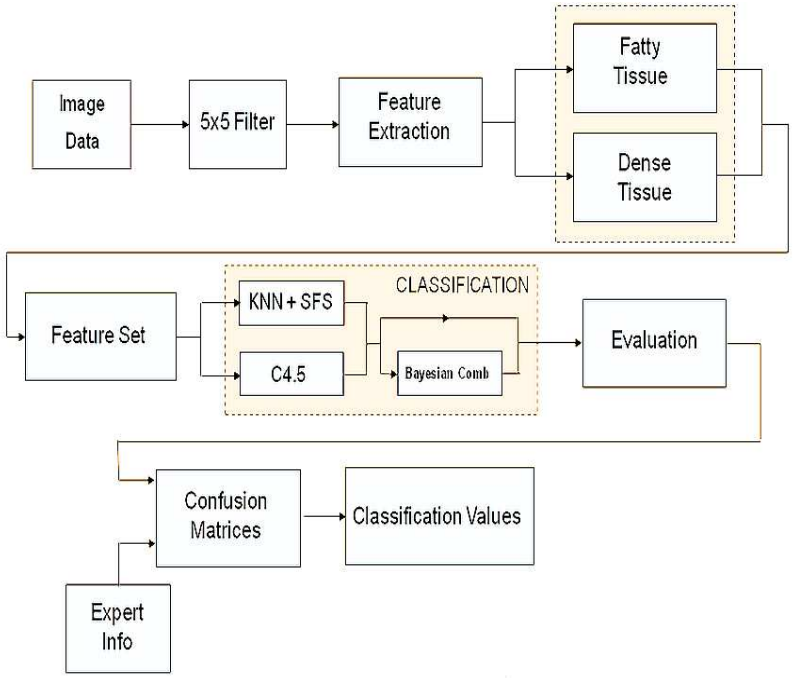

Figure 1: Mammographic Density Classification

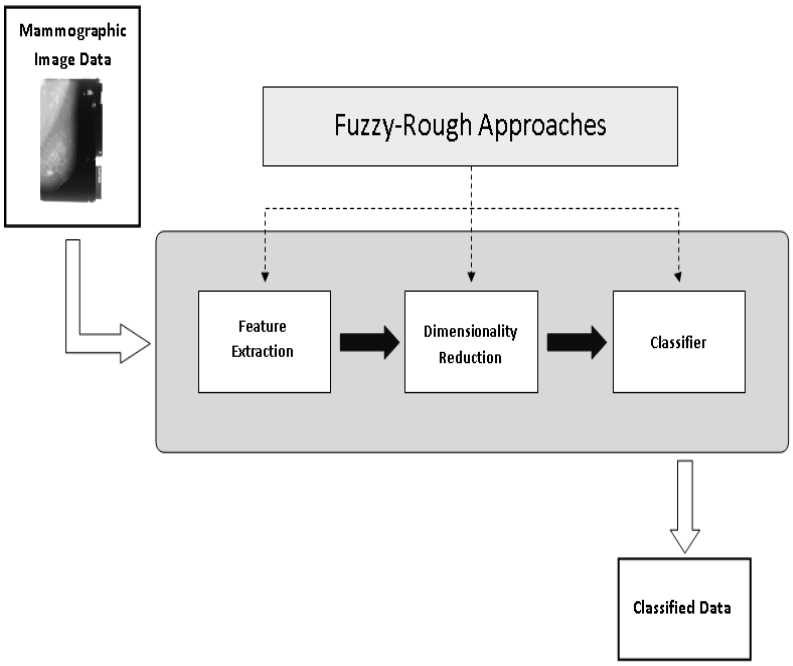

Figure 2: Unified fuzzy-rough framework for mammographic data analysis

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

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

Figure 3: The QUICKREDUCT algorithm

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

(1) $N \leftarrow \mathbb{U}$
(2) $\mu_1(y) \leftarrow 0$, $\mu_2(y) \leftarrow 0$, $Class \leftarrow \emptyset$
(3) $\forall X \in C$
(4) $\mu_{\underline{R_P}X}(y) = \inf_{z \in N} I(\mu_{R_P}(y,z), \mu_X(z))$
(5) $\mu_{\overline{R_P}X}(y) = \sup_{z \in N} T(\mu_{R_P}(y,z), \mu_X(z))$
(6) **if** $(\mu_{\underline{R_P}X}(y) \geq \mu_1(y)$ && $\mu_{\overline{R_P}X}(y) \geq \mu_2(y))$
(7) $Class \leftarrow X$
(8) $\mu_1(y) \leftarrow \mu_{\underline{R_P}X}(y)$, $\mu_2(y) \leftarrow \mu_{\overline{R_P}X}(y)$
(9) **output** *Class*

Figure 4: The FRNN algorithm

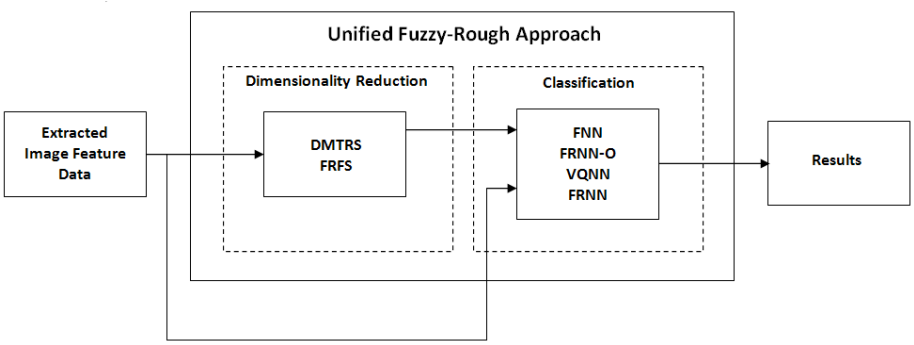

Figure 5: Overview of the experimental evaluation

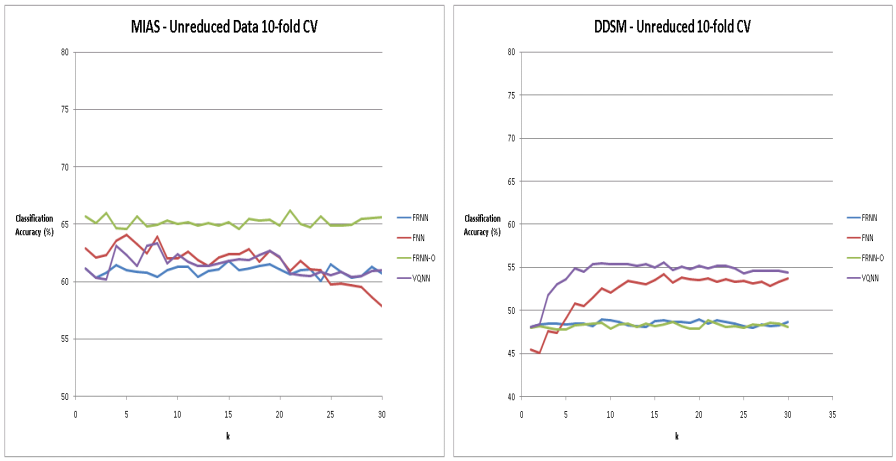

Figure 6: Classification accuracy: Unreduced MIAS and DDSM data

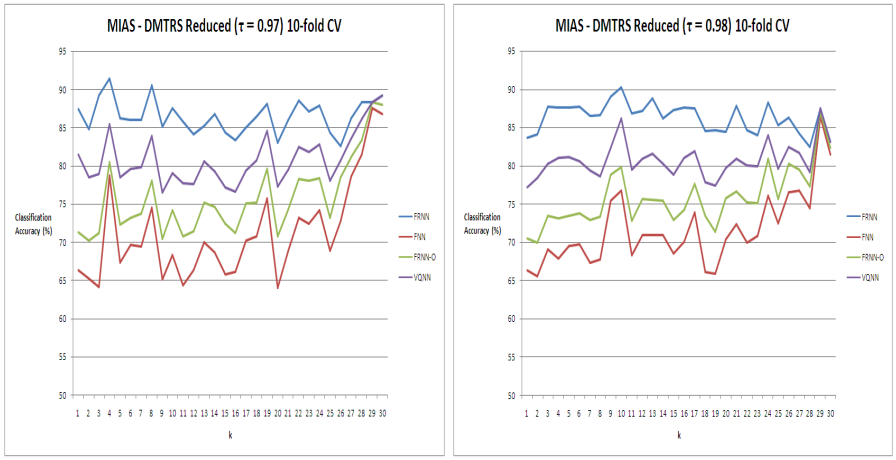

Figure 7: Classification accuracy: DMTRS reduced MIAS data

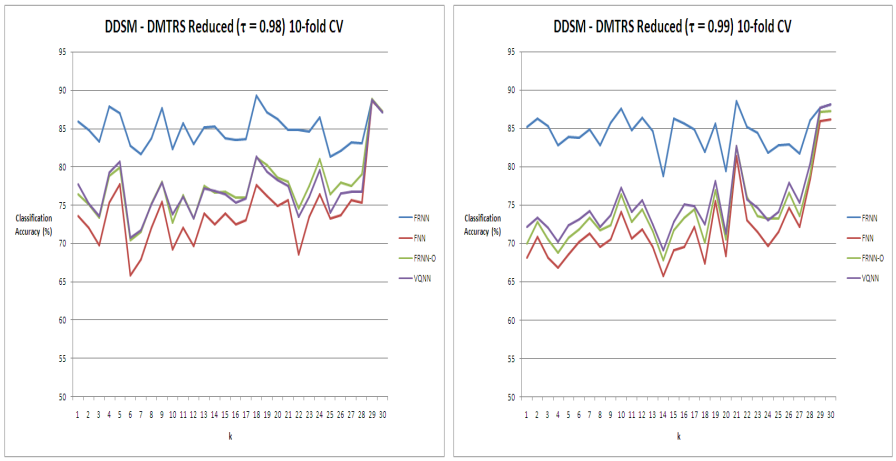

Figure 8: Classification accuracy: DMTRS reduced DDSM data

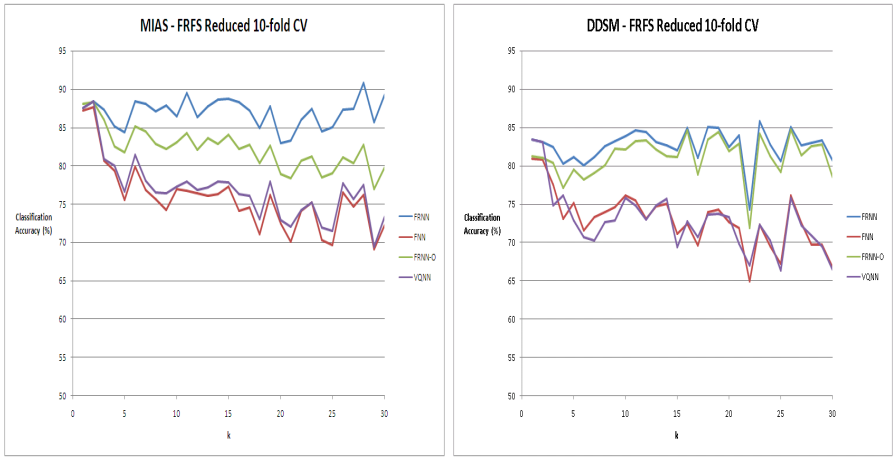

Figure 9: Classification accuracy: FRFS reduced MIAS and DDSM data