Rough and Fuzzy-rough methods for mammographic data analysis

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

Published in:
Proceedings of the 8th Annual UK Workshop on Computational Intelligence

Publication date:
2008

Citation for published version (APA):
Shen, Q., MacParthaláin, N. S., & Jensen, R. (2008). Rough and Fuzzy-rough methods for mammographic data analysis. In *Proceedings of the 8th Annual UK Workshop on Computational Intelligence*
http://hdl.handle.net/2160/661

Rough and fuzzy-rough methods for mammographic data analysis

Neil Mac Parthaláin, Richard Jensen, Qiang Shen
Dept. of Computer Science
Aberystwyth University, Wales
{nsm03,rkj,qqs}@aber.ac.uk

Abstract

The accuracy of methods for the detection of mammographic abnormaility is heavily related to breast tissue characteristics. A breast with high tissue density will have reduced sensitivity in terms of detection. Also, breast tissue density is an important indicator of the risk of development of breast cancer. This paper investigates the application of a number of rough set and fuzzy-rough set techniques to mammographic image data. The aim is to attempt to automate the breast tissue classification procedure based on the consensus data of experts. The results of applying these techniques show that they perform well, achieving high levels of classification accuracy.

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. Furthermore, in the EU and US, breast cancer is recognised as the leading cause of death of women in their 40s. 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 [6, 10]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 abnormalites is therfore highly desirable.

The work in this paper is based on that previously published in [8]. The data, extracted from mammographic images, is used for an approach which considers tissue density with labels for each mammogram object determined using the consensus information of three expert radiologists. There are two datasets which are publicly available from [5] and [14], both of which have four decision classes based on the BIRADS classification protocol [1]. This paper proposes new techniques for classification using rough sets, and fuzzy-rough sets and applied to this mammographic data, incorporating a fuzzy-rough feature selection preprocessing step.

The remainder of the paper is structured as follows. The theoretical background for the fuzzy nearest neighbours classification approaches is detailed in section 2. In section 3, the fuzzy-rough set methodology and its application to both feature selection and the nearest neighbour classification problem is discussed. Section 4 demonstrates the application of the fuzzy-rough nearest neignbours (FRNN) approach to the mammographic data, as well as the experimental setup, and comparative results for each approach. Section 5 concludes the paper.

2 Fuzzy Nearest Neighbours Classification

The K-nearest neighbour (KNN) algorithm [3] is a well-known classification technique that 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 [7].

It allows partial membership of an object to different classes, and also takes into account the relative importance (closeness) of each neighbour with respect to the test instance. However, as correctly argued in [13], 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.

2.1 FNN

The fuzzy K-nearest neighbours algorithm [7] aims to classify test objects based on their similarity to a given number of neighbours and their neighbours' degree of belonging to (crisp or fuzzy) class labels. For the purposes of FNN, the extent to which an unclassified object y belongs to class X is defined as:

$$\mu_X(y) = \sum_{x \in N} \mu_R(x,y)\mu_X(x) \tag{1}$$

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:

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

where $||\cdot||$ denotes the Euclidean norm, and m is a parameter that controls the overall weighting of this fuzzy similarity. The FNN algorithm (figure 1) 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 $O(|\mathbb{U}| + K \cdot |C|)$,

2.2 Fuzzy-rough ownership KNN

Initial attempts to combine the FNN algorithm with concepts from fuzzy rough set theory were presented in [13, 15] (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 uncertainty" (caused by insufficient knowledge, i.e. attributes, about the objects). All training objects influence the

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

(1) $N \leftarrow$ getNearestNeighbours(y,K);
(2) $\forall X \in C$
(3) $\quad \mu_X(y) = \sum_{x \in N} \mu_R(x,y)\mu_X(x)$
(4) **output** $\arg\max_{X \in C}(\mu_X(y))$

Figure 1: The fuzzy KNN algorithm

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. It should be noted that the algorithm does not use fuzzy lower or upper approximations to determine class membership. The fuzzy-rough ownership function was defined as:

$$\tau_X(y) = \frac{\sum_{x \in \mathbb{U}} \mu_R(x,y)\mu_X(x)}{|\mathbb{U}|} \tag{3}$$

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

$$\tau_X(y) = \frac{\sum_{x \in N} \mu_R(x,y)\mu_X(x)}{|N|} \tag{4}$$

where N is the set of object y's K-nearest neighbours. When $K = |\mathbb{U}|$ then the original definition is obtained. The fuzzy similarity is determined by:

$$\mu_R(x,y) = \exp(-\sum_{a \in \mathbb{C}} \kappa_a (a(y) - a(x))^{2/(m-1)}) \tag{5}$$

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

$$\kappa_a = \frac{|\mathbb{U}|}{2\sum_{x \in \mathbb{U}} ||a(y) - a(x)||^{2/(m-1)}} \tag{6}$$

The algorithm can be seen in figure 2. Initially, the 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 equation (3). Finally, when all training objects have been considered, the algorithm outputs the class with highest membership. The complexity of the algorithm is $O(|\mathbb{C}||\mathbb{U}| + |\mathbb{U}| \cdot (|\mathbb{C}| + |C|))$. To obtain

the K-nearest neighbours version of this algorithm, line (4) should be replaced with $N \leftarrow$ getNearestNeighbours(y,K). The method still requires a choice of parameter m, which plays a similar role to that in FNN.

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

(1) $\quad \forall a \in \mathbb{C}$
(2) $\quad\quad \kappa_a = |\mathbb{U}|/2 \sum_{x \in \mathbb{U}} ||a(y) - a(x)||^{2/(m-1)}$
(3) $\quad N \leftarrow |\mathbb{U}|$
(4) $\quad \forall X \in C,\ \tau_X(y) = 0$
(5) $\quad \forall x \in N$
(6) $\quad\quad d = \sum_{a \in \mathbb{C}} \kappa_a (a(y) - a(x))^2$
(7) $\quad\quad \forall X \in C$
(8) $\quad\quad\quad \tau_X(y) += \frac{\mu_X(x) \cdot \exp(-d^{1/(m-1)})}{|N|}$
(9) $\quad$ **output** $\arg\max_{X \in C} \tau_X(y)$

Figure 2: The fuzzy-rough ownership NN algorithm

3 Fuzzy-Rough Set Theory

Over the past ten years, rough set theory (RST) has become a topic of great interest to researchers and has been applied to many domains [9, 4]. Given a dataset with discretized attribute values, it is possible to find a subset (termed a *reduct*) of the original attributes using rough set theory that are the most informative; all other attributes can be removed from the dataset with minimal information loss.

However, quite often attributes values are both crisp and *real-valued*, and this is where traditional rough set theory encounters a problem. It is not possible in the original theory to say whether two attribute values are similar and to what extent they are the same; for example, two close values may only differ as a result of noise, but RST considers them as different as two values of a dissimilar magnitude. It is, therefore desirable to develop techniques which provide a method for knowledge modelling of crisp and real-value attribute datasets which utilise the extent to which values are similar. This can be achieved through the use of *fuzzy-rough* sets. 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

3.1 Fuzzy-Rough Feature Selection (FRFS)

Definitions for the fuzzy lower and upper approximations can be found in [12], 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 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:

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

$$\mu_{R_a}(x,y) = \exp(-\frac{(a(x) - a(y))^2}{2{\sigma_a}^2}) \quad (11)$$

$$\mu_{R_a}(x,y) = \max(\min(\frac{(a(y) - (a(x) - \sigma_a))}{(a(x) - (a(x) - \sigma_a))}, \frac{((a(x) + \sigma_a) - a(y))}{((a(x) + \sigma_a) - a(x))}, 0) \quad (12)$$

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 [4] can be defined as:

$$\mu_{POS_{R_P}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{R_P}X}(x) \quad (13)$$

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

$$\gamma'_P(\mathbb{D}) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_{R_P}(\mathbb{D})}(x)}{|\mathbb{U}|} \quad (14)$$

A fuzzy-rough reduct R can be defined as a minimal 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 greedy hill-climbing algorithm can be constructed that uses equation (14) to gauge subset quality. In [4], it has been shown that the dependency function is monotonic and that fuzzy discernibility matrices may also be used to discover reducts.

3.2 Fuzzy-Rough Nearest Neighbours

To perform classification, the algorithm shown in figure 3 is used. The rationale behind the algorithm 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.

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. The complexity of the algorithm is $O(|C| \cdot (2|\mathbb{U}|))$. Although K is not required, it can be incorporated into the algorithm by replacing line (2) with "$N \leftarrow$ getNearestNeighbours(y,K)". As $\mu_{R_P}(x, y)$ gets smaller, x tends to have only have a minor influence on $\mu_{\underline{R_P}X}(y)$ and $\mu_{\overline{R_P}X}(y)$.

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) $\quad \mu_{\underline{R_P}X}(y) = \inf_{z \in N} I(\mu_{R_P}(y, z), \mu_X(z))$
(5) $\quad \mu_{\overline{R_P}X}(y) = \sup_{z \in N} T(\mu_{R_P}(y, z), \mu_X(z))$
(6) $\quad$ **if** $(\mu_{\underline{R_P}X}(y) \geq \mu_1(y)$ && $\mu_{\overline{R_P}X}(y) \geq \mu_2(y))$
(7) $\quad\quad Class \leftarrow X$
(8) $\quad\quad \mu_1(y) \leftarrow \mu_{\underline{R_P}X}(y)$, $\mu_2(y) \leftarrow \mu_{\overline{R_P}X}(y)$
(9) **output** $Class$

Figure 3: The FRNN algorithm

3.3 Vaguely Quantified Rough Sets (VQRS)

Equations (7) and (8) 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 definitions are recovered. Note in particular how the inf and sup operations play the same role as the $\forall$ and $\exists$ quantifiers, and how a change in a single element can thus have a large impact on (7) and (8). 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 [2] proposed to soften the universal and existential quantifier by means of vague quantifiers like *most* and *some*. Mathematically, the vague quantifiers were modelled in terms of Zadeh's notion of a regularly increasing fuzzy quantifier Q: an increasing $[0, 1] \rightarrow [0, 1]$ 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 $[0, 1]$,

$$Q_{(\alpha,\beta)}(x) = \begin{cases} 0, & x \leq \alpha \\ \frac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \alpha \leq x \leq \frac{\alpha+\beta}{2} \\ 1 - \frac{2(x-\beta)^2}{(\beta-\alpha)^2}, & \frac{\alpha+\beta}{2} \leq x \leq \beta \\ 1, & \beta \leq x \end{cases} \tag{15}$$

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

$$\mu^{Q_u}_{\underline{R_P}X}(y) = Q_u\left(\frac{|R_P y \cap X|}{|R_P y|}\right) \tag{16}$$

$$\mu^{Q_l}_{\overline{R_P}X}(y) = Q_l\left(\frac{|R_P y \cap X|}{|R_P y|}\right) \tag{17}$$

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 figure 3 can be adapted to perform VQRS-based nearest neighbours classification by replacing $\mu_{\underline{R_P}X}(y)$ and $\mu_{\overline{R_P}X}(y)$ with $\mu^{Q_u}_{\underline{R_P}X}(y)$ and $\mu^{Q_l}_{\overline{R_P}X}(y)$.

4 Application

The area of mammographic analysis is an important area for research. Much of the mammographic screening and analysis of patients can be automated, but expert opinion is still required especially when the breast tissue is very dense. Previous work has shown promising results in classifying mammographic images in [8], however improvement of performance is desirable

such that automatic diagnosis can be carried out with a high level of confidence.

4.1 Data

The are two datasets considered in this paper, and both available in the public domain: the Mammographic Image Analysis Society database (MIAS) [14], and the Digital Datebase of Screening Mammography (DDSM) [5]. The MIAS dataset is composed of Medio-Lateral-Oblique (MLO) left and right mammograms from 161 women (322 objects), and has 281 features. The DDSM dataset has 832 mammograms (objects) and again 281 features. The data has been labelled using the consensus of 3 experts to define which mammograms belong to each of the four BIRADS [1] classes. Given the relatively high dimensionality of the MIAS and DDSM data, and the fact that this data has been extracted from images, the application of feature selection would have many positive benefits. In particular, the removal of irrelevant, redundant and noisy features which may otherwise adversly affect classifier learners.

4.2 Experimental Setup

The value of K is initialised as 30 and then decremented by 1 each time, resulting in 30 experiments for each dataset. For each choice of parameter K, 10× 10-fold cross-validation is performed. Note that this parameter is essential only for FNN. For FNN and FRNN-O, m is set to 2. For the new approaches, the fuzzy relation given in equation (10) was chosen. 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. The VQNN approach was implemented using $Q_l = Q_{(0.1,0.6)}$ and $Q_u = Q_{(0.2,1.0)}$, according to the general formula in equation (15).

For the FRFS preprocessor, the fuzzy similarity given in equation 12 was used with the Łukasiewicz t-norm ($\max(x + y - 1, 0)$) and the Łukasiewicz fuzzy implicator ($\min(1 - x + y, 1)$). These have been shown to work particularly well for fuzzy-rough feature selection [4].

4.3 Dimensionality Reduction

Applying FRFS as a preprocessor results in only 6 and 7 features being chosen respectively for both MIAS and DDSM datasets. This is a large decrease in dimensionality from the original 281 features for these datasets.

4.4 Comparative Investigation

For the unreduced DDSM dataset (figure 4) the VQNN method performs significantly better than the other methods, but for the unreduced MIAS dataset all methods perform similarly with FRNN-O showing a small advantage interms of accuracy. However, what is most clearly demonstrated in figure 5 is the value of employing feature selection, as all methods show a significant increase in classification accuracy. Here it can be seen that the new 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.

In [8], the best classification accuracies obtained were 77% for the MIAS dataset and 86% for the DDSM dataset using leave-one-out-cross-validation (LOOCV). As can be seen from the results in this paper, the fuzzy-rough methods greatly outperform these methods when using a fuzzy-rough feature selection preprocessing step despite employing 10-fold cross validation.

5 Conclusion

This paper introduced two new FRNN classification approaches and demonstrated how they could be applied to the analysis of mammographic data. The value of the application of FS techniques has also been highlighted and is apparent in the large improvement of classification accuracy in general for all methods.

Areas for future work include the direct comparison of the approaches proposed here with the kNN, C4.5, and Bayesian methods used in [8]. Also, it would be interesting to apply an unsupervised FS approach to the unlabelled MIAS and DDSM data and compare the classification results with those of the FS approach used here.

Acknowledgements

The authors would like to thank Dr. Reyer Zwiggelaar for his assistance and for providing the data used in this paper, and also Dr. Chris Cornelis.

References

[1] American College of Radiology. Illustrated Breast Imaging Reporting and Data System BI-RADS. 3rd Edition, American College of Radiology, 1998.

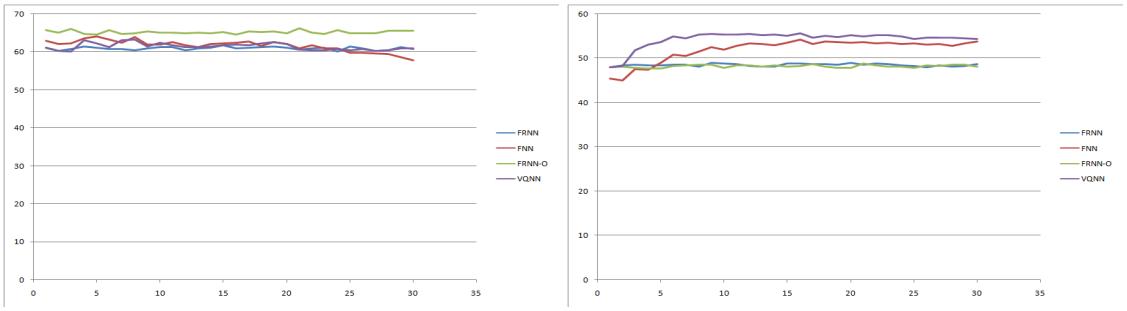

Figure 4: Classification accuracy: Unreduced MIAS and DDSM data (respectively) for the four methods and different values of K(x axis)

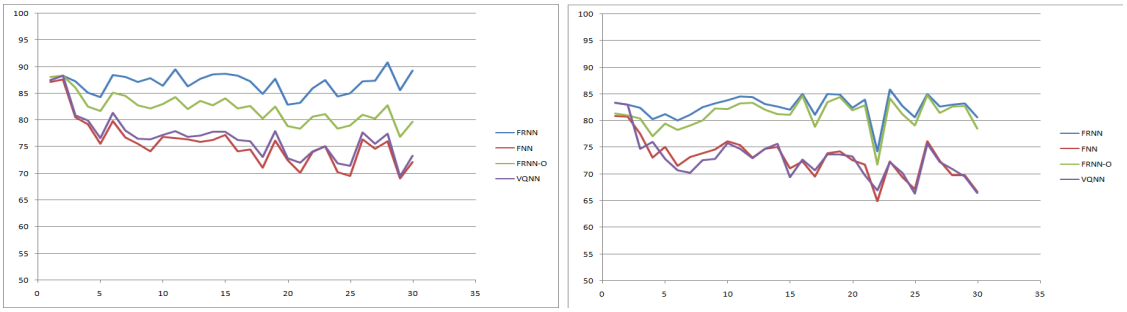

Figure 5: Classification accuracy: Reduced MIAS and DDSM data for the four methods and different values of K (x axis)

[2] C. Cornelis, M. De Cock, A.M. Radzikowska. Vaguely Quantified Rough Sets. Proceedings of 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC2007), Lecture Notes in Artificial Intelligence 4482, pp. 87–94, 2007.

[3] R. Duda and P. Hart. Pattern Classification and Scene Analysis. Wiley, New York, 1973.

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

[5] 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.

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

[7] 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.

[8] 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, Issue 1, pp. 55–65, 2008.

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

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

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

[12] 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.

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

[14] 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.

[15] 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.