Asian Journal of Information Technology 15 (22): 4438-4444, 2016
ISSN: 1682-3915

Rough Set Based Approach for Multiclass Breast Tissue Classification

[1]V.P. Sumathi, [2]K. Kousalya and [1]V. Vanitha
[1]Department of Computer Science and Engineering Kumaraguru College of Technology,
Coimbatore, Tamil Nadu, India
[2]Department of Computer Science and Engineering,
Kongu Engineering College Perundurai, Tamil Nadu, Erode, India

Abstract: Breast cancer is the most common diseases among women and the detection of breast cancer for younger women at an earlier stage can save lives. To identify breast cancer most of the existing work is based on mammography images and has achieved higher accuracy. But the usual mammography test is not suitable and it is not recommended for younger women. Hence, the Electrical Impedance Spectrum (EIS) test for breast cancer detection is more suitable for younger women, the proposed work make use of EIS data to detect breast cancer. The proposed research is based on rough set due to lot of uncertainty in breast tissue data set. The result of classification accuracy and error rate is evaluated. This result indicates that the proposed rough set based approach achieves average accuracy of 65% in six class problem, 80% in four class problem and 85% in two class problem. The two decision classes acquired >80% accuracy using fivefold validation.

Key words: Breast tissue classification, rough set, multiclass classification, rule induction

INTRODUCTION

The risk of developing breast cancer at a young age is increasing due to bad food habits. It is necessary for early detection and identification of young women having breast cancer at the time of early screening. The testing methods like mammograms, Magnetic Resonance Imaging (MRI) or Electrical Impedance Scanning (EIS) are used to detect cancer. Mammography is not recommended routinely for women younger than 40 year. Also it gives less accurate results in women with dense breasts. Accuracy of detection of breast cancer using MRI scan is found to be less. EIS is suitable for younger women and it detects cancer at an earlier stage. Analyzing the data set produced by EIS is important to maximize accuracy and minimize error rate in detection of breast cancer. The medical society is expecting an efficient technique to classify EIS data that can improve the breast cancer treatments for younger women. Most of the analysis works in the existing literatures were based on mammography images. The accuracy of the classification system using mammography images is relatively low for younger women. Usually medical data is having vagueness and uncertainty. Pawlak (1991) initiated new theory called rough set for handling dataset with more uncertainty. In this research, roughest based analysis of breast tissue data is proposed. Breast cancer has multiple stages and identifying cancer stage at an earlier stage is important for giving proper treatment. The cancer stages are indentified by analyzing breast tissue features. For analyzing breast tissue data set rough set based multiclass classification using rule induction technique is proposed. The potential advantage of this breast tissue classification could reduce the screening cost and time to receive the test reports. For classification purpose the breast tissue dataset from University of California Irvine (UCI) repository is taken and it involves six different stages of cancer. Electrical Impedance measurements of freshly excised breast tissue were made at seven differet frequencies. These measurements plotted in the (real, imaginary) plane constitute the impedance spectrum from where the breast tissue features were computed. 120 spectra are collected in excised tissue samples. To classify these data, rough set based classification is proposed and accuracy, error rates are computed. The obtained accuracy is higher than the existing work. The salient features of the proposed work are:

- Rough set based discretization and feature selection in multi class classification problem
- Rough set based rule induction to classify six types of breast tissue
- Implementation of both binary class and multiclass classifiers

Corresponding Author: V.P. Sumath, Department of Computer Science and Engineering Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India

Literature review: The literature review presented in this section is divided into two main parts mammography based breast tissue classification; EIS based classification. In mammography images are used for features extraction. In EIS spectrum were used for feature extraction. There is lots of work based on mammography. Only few works are based on EIS dataset. Vaidehi and Subashini (2015) did the task of identifying the stages of breast tissue and constructing high accurate classification model of cancer from mammography using k-Nearest Neighbors (k-NN) classifier. The obtained accuracy was 91.16%. Due to Breast tissue mammography images high accuracy was scored. Oliver *et al.*(2008) breast tissue types identified using texture and morphological feature of Breast Imaging Reporting and Data System (BIRADS) categories were fed to the C4.5 decision tree, Sequential Forward Selection (SFS) combined with k-NN for classification. They achieved 86% of accuracy. Sheshadri and Kandaswamy (2007) proposed histogram based breast tissue classification using statistical features and achieved 80% of accuracy. Liasis *et al.* (2012) focused on the Scale Invariant Feature Transform (SIFT), Local Binary Patterns (LBP) and textton histogram for Support Vector Machine (SVM) classifier and obtained 93.54% of accuracy. Mustra *et al.* (2012) that considered mini-MIAS database that contains different breast tissue classes based on tissue density. Accuracy of different decision classes was calculated and its vary from 64-91%. Virmani *et al.*, (2016) have chosen medical histogram as input, Principle Component Analysis (PCA) and SVM classifier for medical histograms classification and obtained 94.4% of accuracy.

Normally mammography test is not recommended for younger women. EIS is used to extract the breast tissue features from younger women. By analyzing this data set, identification of cancer developed in younger women at an earlier stage.

The increased density of breast tissue may lead to abnormality. Compare to mammography, EIS is better for dense breasts. Few literatures based on EIS data set were published. Silva *et al.* (2000) have proposed rule based approach to classify the breast tissue dataset. The three features were selected for classification. The overall efficiency was 92% with carcinoma discrimination was greater than 86%. The classification result was very good for training dataset but testing dataset was not classified effectively. Daliri (2015) were used rank based feature selection and Extreme Learning Machine (ELM) combined with SVM for classification and obtained 80% accuracy in two class problem.

MATERIALS AND METHODS

Rough set preliminaries in multiclass classification system: An information system is a pair I = (U, A) where U is a non-empty, finite set called the universe and A is a non-empty, finite set of attributes are called conditions attributes. The Decision Table (DT) is also called an information system and is of the form DT = (U, A ∪ {d}) where d ∉ A is a distinguished attribute called decision classes. Usually the DT contains vagueness and uncertainty data. For handling these types of data the rough set theory has been introduced by Pawlak and Skowron (2007). Rough set is used to identify relationship that would not be found by statistical methods. This relation namely indiscernbility relation which is a binary equivalence relation showing the relation between two instances.

Table 1: Rough set theory notations

Notation	Meaning
U	Universal set
A	Conditional attributes
d	Decision classes various from 1-r
r	Total number of decision classes
X_i	Instances belongs to decision class i
B	Subset of A
$IND_A(B)$	Indiscernibiltiy relation with respect to B
$R_B(x,y)$	Indiscernibility relation exist between the instances x,y based on the attributes in B
$POS_B(d)$	Instances certainly classified as decision class d
$BND_B(d)$	Instances possibly but not certainly classified
ϕ_A, ϕ_B	Generalized decision with respect to A and B
V_d	Values with respect to decision class d
Card	Cardinality various from 0-1
m	Decision rule
$\mu_A(m)$	Coefficient of consistency on rule m
Supp(m)	Number of instances supporting rule m
Match(m)	Number of instances matches with rule m

The various notation used in rough set theory is given in Table 1. In the proposed work the rough set theories applied in multiclass problem. In machine learning, multiclass classification means the problem of classifying instances into one or more decision classes. The multiclass classification problems contain r number of decision classes, X_1.....X_r are subset of instances belong to decision classes d. B is subset of conditional attributes A then the set $BX_1 \cup BX_2$.....BX_r is called the positive region of A denoted by $POS_B(d)$. The function ϕ_B:U->$BND_B(V_d)$, called the B generalized decision of A, by:

$$\phi_B(x) = \{V \in V_d : \exists x^1 \in U(x^1 IND_A(B)_x \text{ and } d(x) = V)\} \quad (1)$$

A generalized decision ϕ_A where A is consistent $Card(\phi_A(x)) = 1$ for any x∈U otherwise A is inconsistent. If A is consistent $POS_A(d) = U$ and m is decision rule in A. The coefficient of consistency of the rule m, $\mu_A(m) = Supp_A(m)/Match_A(m)$ if A is consistent $\mu_A(m) = 1$. The proposed research consists of five steps namely discretization, feature selection, rule induction, rule selection, prediction. The various steps involved in breast tissue classification is shown in Fig. 1.

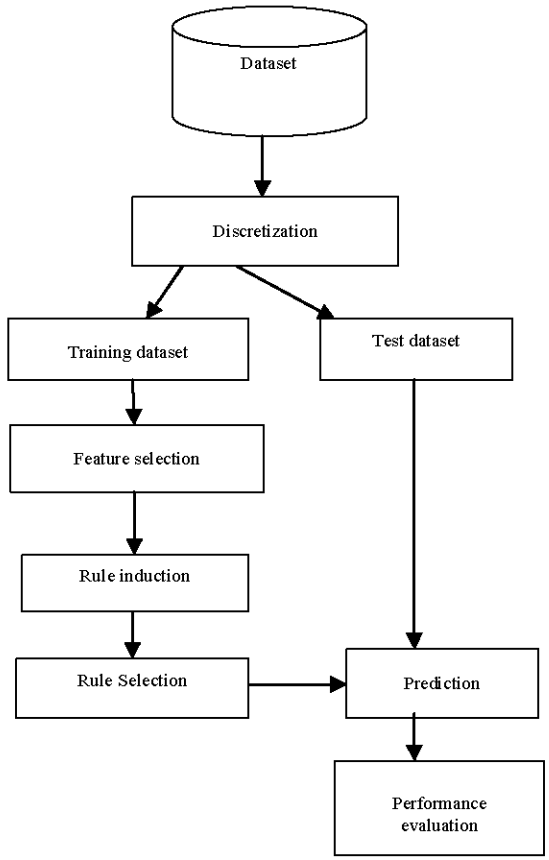

Fig. 1: Steps involved in rough set based data classification

The rough set based algorithms are implemented in every step. The local discretization heuristic algorithm, super reduction algorithm, LEM2 rule induction algorithm, Laplace value based rule selection are implemented in the proposed system

Discretization: Discretization is a popular approach for handling numeric attributes in machine learning if the attributes are both nominal (values that does not possess any order among them) and continuous (values with order among the values). Most of the classifiers require discretized data for classification. The preprocessing of continuous data into discretized data is important step before data being used for rule induction approaches. Nguyen (2001) stated the distribution of continuous data into discrete ranges may affect the accuracy of classification. Local discretization has been used for classification of breast tissue data set. This method carries out discretization during decision tree building process. They produce partitions that are applied to localized regions of the instance space. The partitions are called cut values.

In this system a_i indicates conditional attribute $i \in A$, c_k denotes the k^{th} cut of attributes a_i. The set of all cuts corresponding to a_i is (a_{i1},c_1), (a_{i2},c_2), ……(a_{ik},c_k) stored in P. A is inconsistent if and only if for any pair of instances x, y ∈ U and decision class of x not equal to decision class of y. The objects x and y are discernible by attributes from A then there exists (a_{ij},c_j) ? P discerning x and y. The problem of finding the optimal set of cuts P in a given decision table DT is NP-hard. Semi-optimal solution can be identified using some approximation algorithms. Divide and conquer strategy to determine the best cut $C_{best} \in \{c_1, c_2, c_3, \ldots c_n\}$ with respect to quality function where n is number of cuts.

Algorithm 1 to find semi-optimal cut:

Algorithm : Semi-optimal cut

Step1:consider attribute a, divide the attribute value range into number of cuts. The set of cuts corresponding to attribute a is input $c_a = \{c_1, c_2, c_3, \ldots. c_N\}$

Step2: compute the quality of each cut in the given input

Step3:Based on the quality, the cuts in the order like increasing from $c_1, c_2, \ldots c_{min}$ and decreasing from $c_{max}, c_{max+1}, \ldots. c_N$.

Step4:choose the c_{best} lies between $c_{min}, c_{min+1} \ldots c_{max}$. The input c_a become $\{c_{min}, c_{min+1}, \ldots C_{max}\}$

Step5: repeat the steps from one to four until $c_{min} < c_{max}$

Step6:choose the semi-optimal cut c_{best} lies between c_{min}, c_{max}.

First divide the attribute value range in to specified number of N intervals. By using approximate discernible measure to predict the interval which most probably contains the best cut. The process is repeated until the considered interval consists of one cut then the best cut is chosen between all visited cuts.

Feature selection: It is a process of finding a subset of attributes that represents the same information as the complete feature set. In other words the importance of feature selection is to identify significant attributes and to eliminate the trivial attributes (Wroblewski, 2001). An attribute a∈B, B is subset of A can be regarded as trivial attributes in B if indiscernibility equivalence relation $R_B = R_{B/[a]}$ otherwise a is called significant attributes in B. A super reducion is a set of attributes B subset of A such that $R_B = R_A$ where R_B and R_A are indiscernibility relations defined by B and A respectively. In super reduction algorithm (Janusz and Stawicki, 2011) the feature selection is based on degree of consistency (D_B). It is calculated using Eq. 1:

$$D_B = |POS_B| / |U|$$

Algorithm 2 feature selection algorithm:

Algorithm: Super reduction

//Input : A decision table DT = (U,A∪ {d})

//Output: A Set of selected features SR

SR = { };

Repeat
T= SR;
For each x?(A-SR) do
Compute D_T and $D_{SR\cup\{x\}}$
If $D_{SR\cup\{x\}} > D_T$ then T= SR∪{x}
SR= T
End
Until $D_{SR} = = D_A$

The Decision Table (DT) is called consistent if $D_{DT} = 1$. The super reduction algorithm selects the features which are having high degree of consistency than remaining unselected attributes. The selected set of attributes is used as input to the rule induction technique.

Rule induction: Knowledge representations are in many different forms. Production rules are the most popular knowledge representation. The general structure of rules are IF ... THEN where IF predecessor THEN successor. The advantage of this method is easy to understand and manipulate. In RST a rules derived based on decision Table DT is denoted by:

IF Set (A, Va) THEN d = w

Where:
A = Condition attributes in DT
V_a = Value of the attribute a and w denotes value of decision class d

The important properties of decision rules are completeness and consistency three different types of rule induction approaches which are:

- Minimum-Generates smallest number of decision rules
- Exhaustive-generates all possible decision rules
- Satisfactory-generates rules that meet the predefined requirements

The learning systems are always handling uncertainty and noise data. Two main reasons for uncertainties incomplete, conflict. Rough set theory is well suited to deal with inconsistent data. Based on RST two set of rules are induced namely certain and possible. Certain rules are categorical and possible rules may deals with uncertainty (Stefanowski, 1998). The system producing certain and possible rules is called Learning from Example based on Roughset (LERS). Rule induction algorithms are two types one global induction which covers all attribute values is called Learning from Examples Module version 1 (LEM1) and local induction consider set attribute-value pair Learning from Examples Module version 2 (LEM2). The algorithm LEM2 depends on the idea of lower and upper approximation of a concept represented by a decision value pair (d,w). (attribute, value)-> (decision, value). The algorithm LEM2 (Grzymala, 1992) is based on computing a single local covering for each of the concepts from decision Table.

Rule selection: The decision rules generalized by using rough set methods cannot be acceptable. This situation occurs when the number of examples supporting the decision rule is small. To overcome this problem the strength of the rule is computed based on product of cardinality of support of a rule and length of the rule. The Support (S) means total number of transaction that hold the set of conditional attribute values $x_1,x_2,...x_m$ and decision class y_j is given in Eq. 2. The total number of selected attributes is m and j indicates decision class from which the instance belongs. The Confidence (C) is calculated by considering how often decision class y_j appears in instances that contain conditional attribute values $x_1,x_2,...x_m$ as given in Eq. 3:

$$S = \sigma\ (x_1, x_2\ \text{and}\ x_m, y_j\ )/|T| \quad (2)$$

$$C = \sigma\ (x_1, x_2\ \text{and}\ x_m, y_j\ )/\sigma(x_1, x_2, \text{and}\ x_m) \quad (3)$$

where, t-total number of transactions the laplace used to estimate the accuracy of the rules, the expected accuracy of the rules is calculated using Eq. 4:

$$\text{Laplace Accuracy} = (n_c+1)/(n_{tot}+k) \quad (4)$$

Where:
n_c = No. of examples belongs to 'c' class
k = Total number of decision classes in a DT
$n_{tot=}$ = Total number of examples satisfying the rule's body

LEM2 algorithm is used to generate list all possible rules. The support and confidence is compared for each rule. Then rules reduction is carried out based on the value of support and confidence. The rules are having very less support and confidence values are removed from the generated rule list. The remaining rules are going to be used for prediction.

Prediction: The reduced rule set contains the best krules for each decision class. The prediction is being done based on multiple rules because one rule cannot be sufficient for predicting the correct target class. Each target class has different number of rules. The rules with lower accuracy are not used for prediction. The test data set are predicted and the accuracy of the prediction is calculated.

RESULTS AND DISCUSSION

The system is implemented with rough set theory using R language (Riza *et al*., 2014) and conducted an extensive performance analysis to evaluate the accuracy of multiclass classification system. The two parameter accuracy and error rate is used to evaluate the performance of the system. The accuracy and error rate is computed using Eq. 5 and 6. The percentage of accuracy and error rate is calculated by comparing the predicted values and real values of the test dataset. The sum of accuracy and error is equal to hundred percent:

$$\text{Accuracy (\%)} = 100*\frac{\text{Total number of transaction correctly predicted}}{\text{Total transactions}} \quad (5)$$

$$\text{Error (\%)} = 100\times\frac{\text{Total number of transactions wrongly predicted}}{\text{Total transactions}} \quad (6)$$

The accuracy and error rate percentage are calculated by considering various number of decision classes. The dataset is taken from UCI machine learning repository (Lichman, 2013) that contains 106 instances of breast tissue data. This dataset contains six decision classes namely car (carcinoma), fad (fibro-adenoma), mas (mastopathy), gla (glandular), con (connective), adi (adipose). Three classes (car, fad, mas) indicates cancer condition, another three classes (gla, con, adi) indicates non cancer condition. The dataset can be used for predicting the classification of six decision classes. The attribute information is given in Table 2.

In four class problem fad, mas, gla are merged together and considered as a single class whose discrimination is not important. In two class problem the tissue types are grouped in two decision class namely normal tissue and pathological tissue. The calculated accuracy of the classification is shown in Table 3 where three different subset of features set are considered for breast tissue classification. The graphical representation of classification is shown in Fig. 2 . Each bar represents the different set of features considered for classification. The performance of four class and two class classification is found to be good. Confusion matrix for six, four and two class classification problems are shown in Table 4-6. Diagonal cell in these tables indicates exact prediction of every decision class.

Table 4 maximum accuracy for predicting all six decision classes in five fold validation is 71.43% (15/21) and error rate is 28.57% (6/21).

Table 2: Attribute information

Attribte name	Purpose of the attribute
I0	Impedivity (ohm) at zero frequency
PA500	Phase angle at 500 KHz
HFS	High-frequency slope of phase angle
DA	Impedance distance between spectral ends
AREA	Area under spectrum
A/DA	Area normalized by DA
MAX IP	Maximum of the spectrum
DR	Distance between I0 and real part of the maximum frequency point
P	Length of the spectral curve

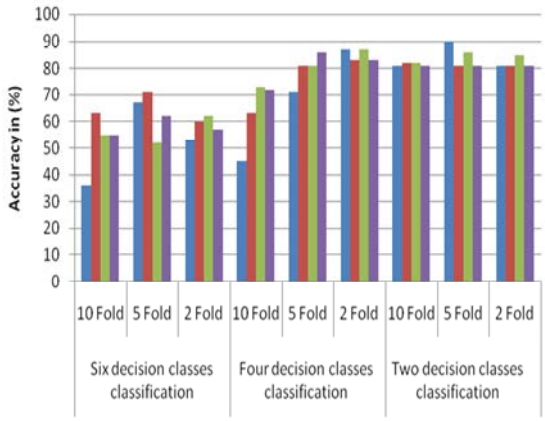

Fig. 2: Algorithm to find Semi-optimal cut

Similarly from Table 5 maximum accuracy for predictions four decision class problem in five fold validation is 80.95% (17/21) and error rate is 19.05% (4/21). From Table 6 the maximum accuracy for predicting two decision class in five fold validation is 90.47% (19/21) and error rate is 9.53% (2/21).

The proposed work is compared with three existing approaches. The first approach (Silva *et al*., 2000) proposed classification system based on statistical analysis and rule induction technique. The maximum accuracy obtained was 92%. The dataset used by the researcher for their experiments is not same as the one they donated later in 2010. The second approach proposed a system in which feature selection is based on rank and classification is based on SVM and ELM. In this system the accuracy for two class problem is 80%. The third approach (Daliri, 2015) is similar to second approach. The accuracy obtained by this system was given as high for six class problem. Figure 3 clearly shows that the maximum accuracy of two class classification problem in five fold validation is 90.47% for the proposed approach and it was 80% for approach two.

The breast tissue data set classification accuracy using rough set based proposed classification system is

Table 3: Performance of the system in 6, 4 and 2 class problem

		Accuracy (%)			
No. of decision classes	Validation	All nine features	IO, P, DA, DR, Area	IO, PA500	PA500, P
Six	10 fold	36	63	55	55
	5 fold	67	71	52	62
	2 fold	53	60	62	57
Four	10 fold	45	63	73	72
	5 fold	71	81	81	86
	2 fold	87	83	87	83
Two	10 fold	81	82	82	81
	5 fold	90	81	86	81
	2 fold	81	81	85	81

Table 4: Confusion matrix for six decision classes

Decision classes	Car	Fad	Mas	Gla	Con	Adi	Accuracy (%)
Car	4		1				80.00
Fad			1				0.00
Mas			3			1	75.00
Gla				2		1	66.60
Con					1	2	33.30
Adi						5	100.00

Table 5: Confusion matrix for four decision class

Decision classes	Car	Fad mas gla	Con	Adi	Accuracy (%)
Car	4	1			80.00
Fad					
Mas					
Gla	1	7			87.50
Con			1	2	33.33
Adi				5	100.00

Table 6: Confusion matrix for binary classification

Decision classes	Normal tissue (gla,con,adi)	Pathological tissue (car,fad,mas)	Accuracy (%)
Normal tissue (gla, con, adi)	10	1	90.90
Pathological tissue (car, fad, mas)	1	9	90.00

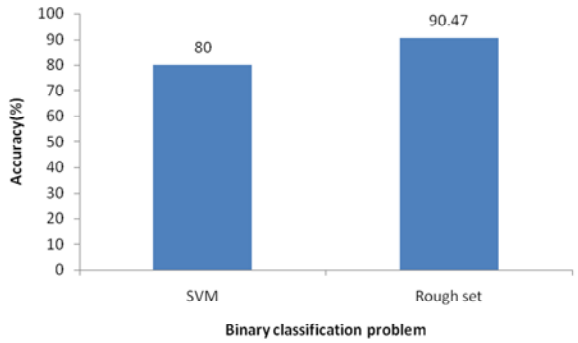

Fig. 3: Performance of multiclass classification

high compared to SVM classification system. Hence, uncertainty is addressed well using rough set.

CONCLUSION

In the proposed system, the rough set based approach is used for multiclass classification problem. The rough set theory is used in every stage of classification process. Finally the accuracy of the classification techniques is evaluated by doing two, five and ten fold validation. The overall accuracy of the system is improved than the previous results. The advantage of this work is different number of decision classes classification are performed. But accuracy six class classifications is not satisfied due to instant nature. Summarizing the result, the maximum accuracy obtained in six class problem is 71.43%, four class problems is 80.95% and two class problem is 90.47% in five fold validation. The error rate also calculated which is minimum in two class problems (9.53%) and maximum in six class problems (28.57%). The rough set based technique improved the accuracy for predicting car (carcinoma) and adi (adipose) decision classes whose classification is almost >80%.

REFERENCES

Daliri, M.R., 2015. Combining extreme learning machines using support vector machines for breast tissue classification. Comput. Methods Biomech. Biomed. Eng., 18: 185-191.

Grzymala, B.J.W., 1992. LERS-A System For Learning From Examples Based on Rough Sets. In: Intelligent Decision Support. Roman, S. (Ed.). Springer Netherlands, Berlin, Germany, ISBN: 978-90-481-4194-4, pp: 3-18.

Janusz, A. and S. Stawicki, 2011. Applications of Approximate Reducts to the Feature Selection Problem. In: Rough Sets and Knowledge Technology. JingTao, Y., S. Ramanna, G. Wang and Z. Suraj (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-642-24424-7, pp: 45-50.

Jossinet, J., 1996. Variability of impedivity in normal and pathological breast tissue. Med. Biol. Eng. Comput., 34: 346-350.

Liasis, G., C. Pattichis and S. Petroudi, 2012. Combination of different texture features for mammographic breast density classification. Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics and Bioengineering (BIBE), November 11-13, 2012, IEEE, Nicosia, Cyprus, ISBN: 978-1-4673-4357-2, pp: 732-737.

Lichman, M., 2013. UCI Machine Learning Repository. University of California, Irvine, California,.

Mustra, M., M. Grgic and K. Delac, 2012. Breast density classification using multiple feature selection. Automatika J. Control Measure. Electron. Comput. Commun., 53: 362-372.

Nguyen, H.S., 2001. On efficient handling of continuous attributes in large data bases. Based Inf., 48: 61-81.

Oliver, A., J. Freixenet, R. Marti, J. Pont, E. Perez, E.R.E. Denton and R. Zwiggelaar, 2008. A novel breast tissue density classification methodology. IEEE Trans. Inform. Technol. Biomed., 12: 55-65.

Pawlak, Z. and A. Skowron, 2007. Rudiments of rough sets. Inform. Sci., 177: 3-27.

Pawlak, Z., 1991. Rough Sets: Theoretical Aspects of Reasoning about Data. 1st Edn., Kluwer Academic Publishers, London, UK., ISBN-13: 9780792314721.

Riza, L.S., A. Janusz, C. Bergmeir, C. Cornelis and F. Herrera et al., 2014. Implementing algorithms of rough set theory and fuzzy rough set theory in the R package roughsets. Inf. Sci., 287: 68-89.

Shen, Q. and A. Chouchoulas, 2000. A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems. Eng. Appl. Artif. Intell., 13: 263-278.

Sheshadri, H.S. and A. Kandaswamy, 2007. Experimental investigation on breast tissue classification based on statistical feature extraction of mammograms. Comput. Med. Imaging Graphics, 31: 46-48.

Silva, D.J.E., D.J.M. Sa and J. Jossinet, 2000. Classification of breast tissue by electrical impedance spectroscopy. Med. Biol. Eng. Comput., 38: 26-30.

Stefanowski, J., 1998. On rough set based approaches to induction of decision rules. Rough Sets Knowl. Discovery, 1: 500-529.

Vaidehi, K. and T.S. Subashini, 2015. Breast tissue characterization using combined K-NN classifier. Indian J. Sci. Technol., 8: 23-26.

Virmani, J., N. Dey and V. Kumar, 2016. PCA-PNN and PCA-SVM Based CAD Systems For Breast Density Classification. In: Applications of Intelligent Optimization in Biology and Medicine. Hassanien, A.E., C. Gorsan and M.F. Tolba (Eds.). Springer International Publishing, Berlin, Germany, ISBN: 978-3-319-21211-1, pp: 159-180.

Wroblewski, J., 2001. Ensembles of classifiers based on approximate reducts. Based Inf., 47: 351-360.