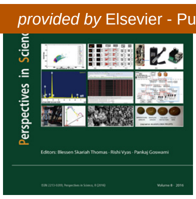
Perspectives in Science (2016) **8**, 247—250

ScienceDirect

journal homepage: www.elsevier.com/pisc



Genetic Fuzzy System (GFS) based wavelet co-occurrence feature selection in mammogram classification for breast cancer diagnosis[☆]

Meenakshi M. Pawar[a,*], **Sanjay N. Talbar**[b]

[a] *Electronics and Telecommunication Engineering, SVERI's College of Engineering, Pandharpur District, Solapur 413304, India*
[b] *Electronics and Telecommunication Engineering, S.G.G.S.I.E. & T, Nanded, India*

Received 2 February 2016; accepted 11 April 2016
Available online 27 April 2016

KEYWORDS
Digital mammogram;
Genetic Fuzzy System algorithm;
Feature selection

Summary Breast cancer is significant health problem diagnosed mostly in women worldwide. Therefore, early detection of breast cancer is performed with the help of digital mammography, which can reduce mortality rate. This paper presents wrapper based feature selection approach for wavelet co-occurrence feature (WCF) using Genetic Fuzzy System (GFS) in mammogram classification problem. The performance of GFS algorithm is explained using mini-MIAS database. WCF features are obtained from detail wavelet coefficients at each level of decomposition of mammogram image. At first level of decomposition, 18 features are applied to GFS algorithm, which selects 5 features with an average classification success rate of 39.64%. Subsequently, at second level it selects 9 features from 36 features and the classification success rate is improved to 56.75%. For third level, 16 features are selected from 54 features and average success rate is improved to 64.98%. Lastly, at fourth level 72 features are applied to GFS, which selects 16 features and thereby increasing average success rate to 89.47%. Hence, GFS algorithm is the effective way of obtaining optimal set of feature in breast cancer diagnosis.

Introduction

Breast cancer is the serious health problem and mostly found in women around 40 ages. New cancer cases are estimated as among women 231,840 and men 2350 cases. ACS also estimated 40,730 deaths due to breast cancer

[☆] This article belongs to the special issue on Engineering and Material Sciences.
[*] Corresponding author. Tel.: +91 9545553666; fax: +91 02186 225082.
E-mail address: minakshee2000@gmail.com (M.M. Pawar).

http://dx.doi.org/10.1016/j.pisc.2016.04.042

during the 2015 (American Cancer Society, 2015). Therefore, breast cancer should be detected in its early stage, which will reduce the number of deaths. Digital mammography is clinically accepted imaging modality worldwide for detection of breast cancer in its early stage (Eltoukhy et al., 2012). Examination of large volume of mammogram may cause extreme tiredness as a result of which radiologist may neglect some important clues. Researchers have demonstrated that 10—30% of the visible cancers on mammograms are overlooked and only 20—30% of biopsy cases are actually resulted as malignant tissues (Dheeba et al., 2014; Eltoukhy et al., 2012). Therefore, Computer aided diagnosis (CAD) system in mammogram classification can work as second reader for radiologist.

Usefulness of wavelet and curvelet transform based feature extraction approach has been studied for mammogram classification (Dhahbi et al., 2015; Beura et al., 2015; Dheeba et al., 2014; Eltoukhy et al., 2012). Methods proposed in (Eltoukhy et al., 2012) are based on signature vector extracted from biggest wavelet coefficients. However, size of signature vector is large and needs more computations. Therefore, it needs to focus on reduced set of features. In view, Genetic algorithm was used for selection of most relevant features (Ramos et al., 2012), presenting an AUC = 0.9. Only 8-selected curvelet moments have been used in (Dhahbi et al., 2015) for malignancy detection and 81% classification accuracy is obtained. Beura et al. (2015) applied *t*-test and *F*-test on feature matrix to obtain relevant features. Aforementioned studies (Dhahbi et al., 2015; Beura et al., 2015; Ramos et al., 2012) have been demonstrated only for classification of breast tissues as benign or malignant.

Fuzzy logic is better for mammogram analysis (Vadive and Surendiran, 2013) and its computational complexity is less as compare to other tools (Pawar, 2006). Therefore, current mammogram classification problem uses feature extraction based on wavelet co-occurrence features (WCF), feature selection for breast abnormality using Genetic Fuzzy System. The rest of paper is described as section ''Genetic Fuzzy system (GFS) based feature selection algorithm'' explains Genetic Fuzzy System (GFS) based feature selection algorithm; section ''Performance analysis of Genetic Fuzzy System'' presents performance analysis of Genetic Fuzzy System whereas conclusion is mentioned finally.

Genetic Fuzzy System (GFS) based feature selection algorithm

The proposed system is developed using preprocessing as region of interest (ROI) selection, feature extraction from wavelet coefficients and feature selection with GFS system.

Preprocessing

Mammogram images from MIAS data set are used for experimentation work as the previous studies (Dhahbi et al., 2015; Beura et al., 2015; Dheeba et al., 2014; Ramos et al., 2012; Eltoukhy et al., 2012) have used it for experimentation purpose. The details about MIAS dataset are given in Table 1. Original mammogram image of 1024 × 1024 pixels is cropped into 128 × 128 pixels image called as ROI image, which reduces background noise.

Table 1 Details about MIAS dataset.

Various cases of breast abnormality	Benign	Malignant	Total
Normal	—	—	27
Circumscribed mass (CIRC)	19	4	23
Microcalcification (CALC)	12	13	25
Spiculated mass (SPIC)	11	8	19
ill-defined mass (MISC)	7	7	14
Asymmetry (ASYM)	6	9	15
Architectural distortion (ARCH)	9	10	19
Total	64	51	142

Feature extraction

Discrete wavelet transform is used to decompose mammogram image using db8 wavelet function. Wavelet coefficients from detail subbands (horizontal, vertical and diagonal) are used to form GLCM matrix ($G(i,j \mid d, \theta)$ where $d = 1$ and $\theta = 0°$). Texture features viz. energy, cluster prominence, cluster shade, sum variance, sum average and entropy are calculated from GLCM matrix and are denoted as wavelet co-occurrence features (WCF). Each wavelet subband gives 6 features and each level gives 18 features. Therefore, total 72 features are obtained from four level of mammogram decomposition.

Genetic Fuzzy System for feature selection

GFS is constructed by combining fuzzy classifier with genetic algorithm as shown in Fig. 1. This section provides brief description about fuzzy classifier and formulation optimization problem for GFS so as to perform feature selection as well as maximization of classification performance.

Fuzzy classifier

Fuzzy classifier is developed using four units viz. Fuzzification, rules, inference engine and defuzzification. Fuzzification unit converts each crisp input (WCF) features into fuzzy sets (F_i) and is specified using membership value. Membership value for each feature is calculated using Gaussian membership function as

$$\mu(f) = e^{-0.5((f-mean)/std.dev.)^2} \quad (1)$$

where, $f = \{f_1, f_2, \ldots, f_m\}^T \in R^m$ represents m input features, mean and std. dev. are computed from features of samples each breast abnormality.

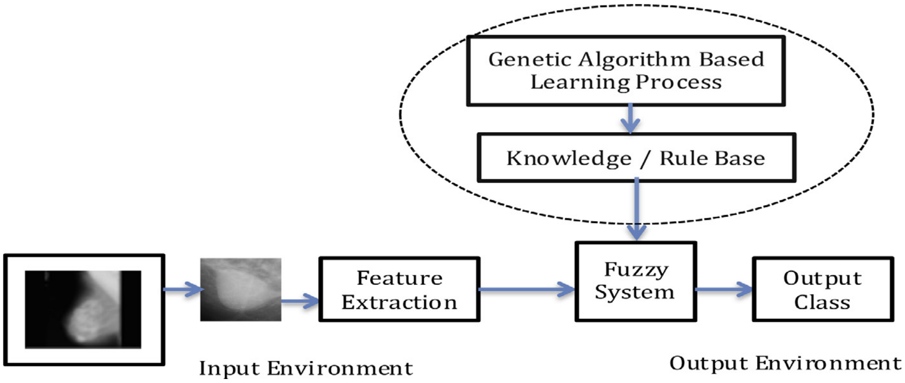


Figure 1 Schematic for GFS.

Rule development and Inference engine

Fuzzy rule for each abnormality is developed using fuzzy sets (F_{im}). Therefore, the ith rule will be viewed as

$$R_i : IF\ f_1\ is\ F_{i1}\ AND\ f_2\ is\ F_{i2}\ AND \ldots f_m\ is\ F_{im}\ THEN\ y = C_i \qquad i = 1, 2, \ldots, L \tag{2}$$

where m is number of WCF features and L represents number of rules or number of output classes ($C = \{C_1, C_2, \ldots, C_L\}^T \in R$) of fuzzy classifier.

Defuzzification

Final step of fuzzy classifier is defuzzification, which uses maximum matching method. Degree of matching of ith rule with kth pattern can be found as $D_k^i = \prod_{p=1}^{m} \mu_{pi}$, where μ_{pi} represents membership value of pth WCF feature in ith rule of fuzzy region. Therefore, the maximum matching of L rules give output class label C_1 as

$$D_k^{\max}(C_i) = \max_i D_k^i (C_i) \tag{3}$$

The classification accuracy (CA) is calculated after defuzzification process as

$$CA = \frac{N_C}{N_T} * 100 \tag{4}$$

where N_C denotes number of correctly classified samples ($y = C_i$) and N_T as the total samples.

Formulation of optimization problem for GFS

For feature selection, Genetic algorithm is used to reduce fitness value β (classification error β) in turn maximizes the

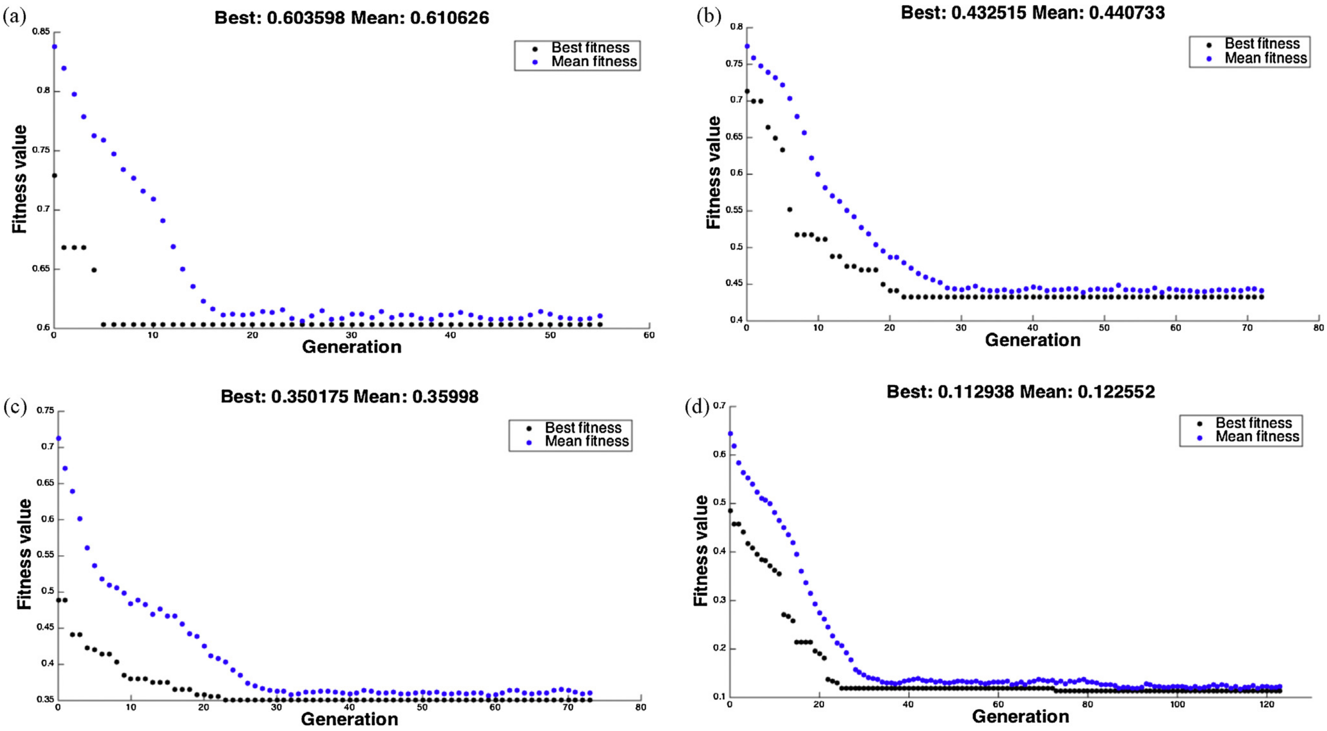


Figure 2 Plot for fitness value versus number of generation using GFS at (a) Level 1, (b) Level 2, (c) Level 3 and (d) Level 4.

Table 2 Number of feature selected at each level of decomposition.

Name of WCF Feature	Number of features selected at each level of decomposition			
	Level 1	Level 2	Level 3	Level 4
Energy	01	01	01	01
Cluster prominence	02	05	05	01
Cluster shade	01	01	04	04
Sum variance	—	01	02	02
Sum average	—	—	03	03
Entropy	01	01	01	05
Total	05	09	16	16

classification accuracy (CA). Thus, the optimization problem is defined as

$$minimize \sum_{i=1}^{L} \beta(i)$$

$$subject\ to: \ \alpha_{il} \in \{0\,1\}$$

$$\mu_{il} = 1, \quad \alpha_{il} = 0;$$

$$\mu_{il} = \mu_{il}, \quad \alpha_{il} = 1 \quad \text{for } l = 1, 2, \ldots, m \tag{5}$$

where, L is the number of rules and μ_{il} membership value of ith rule for lth in fuzzy region.

Performance analysis of Genetic Fuzzy System

The performance of GFS is analysed at each level of decomposition using WCF features as shown in Fig. 2. At Level1, 05 features are selected from 18 features with 39.64% classification accuracy. Subsequently, 36 features from level1 and level2 are applied to GFS, which selects 9 features and classification accuracy is improved to 56.75%. Further, 54 features of level1, 2 & 3 are used for GFS and 16 features are selected with classification accuracy of 64.98%. Finally, 72 features from level1 to level4 are used out of which 16 discriminative features are selected and classification accuracy is improved to 89.47%. Table 2 describes number of feature selected at each level of decomposition. It is observed that levels 3 and 4 select same number of features but its distribution is different. Therefore, we have obtained different classification performance. Classification accuracy from level1 to level4 is represented in Fig. 3 and it is also noted that the classification accuracy is improved for each type of abnormality in level4, for normal 100%, circumscribed mass 80%, calcification as 92.59%, spiculated mass as 94.73%, ill-defined mass 93.33%, architectural distortion 78.94% and asymmetry 86.67%.

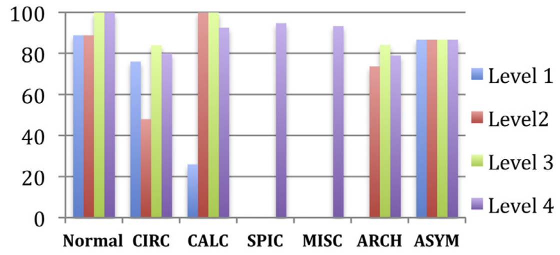


Figure 3 Classification accuracy for Level 1, Level 2, Level 3 and Level4.

Conclusion

The proposed CAD system uses WCF features from four level of decomposition to select optimal feature set and maximize classification accuracy using GFS. The mammogram images are converted into corresponding wavelet coefficients using db8 wavelet function. WCF features were computed from wavelet coefficients of detail sub-bands from mammogram decomposition. Performance of GFS system is demonstrated using mammograms from MIAS database. The highest classification accuracy of 89.47% is achieved with only 16 discriminative features from four level of mammogram decomposition. Hence, it is advantageous to use GFS system for mammogram classification.

References

American Cancer Society, 2015. http://www.cancer.org/Research/CancerFactsFigures/index.

Beura, S., Majhi, B., Dash, R., 2015. Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. Neurocomputing 154, 1—14.

Dhahbi, S., Barhoumi, W., Zagrouba, E., 2015. Breast cancer diagnosis in digitized mammograms using curvelet moments. Comp. Biol. Med. 64, 79—90.

Dheeba, J., Albert Singh, N., Tamil Selvi, S., 2014. Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach. J. Biomed. Inf. 49, 45—52.

Eltoukhy, M.M., Faye, I., Samir, B.B., 2012. A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation. Computers in Biology and Medicine 42, 123—128.

http://peipa.essex.ac.uk/ipa/pix/mias.

Pawar, P.M., (Ph.D. thesis) 2006. Structural Health Monitoring of Composite Helicopter Rotor Blades. Indian Institute of Science, Bangalore, India.

Ramos, R.P., do Nascimento, M.Z., Pereira, D.C., 2012. Texture extraction: an evaluation of ridgelet, wavelet and co-occurrence based methods applied to mammograms. Expert Syst. Appl. 39, 11036—11047.

Vadive, A., Surendiran, B., 2013. A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories. Comp. Biol. Med. 43, 259—267.