International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 3, May-June 2016 **ISSN 2278-6856**

Genetic Association Rule Mining Using Intensity Histogram and GLCM features

[1]Aswini kumar Mohanty, Amalendu Bag[2] Devitosh Acharyar[3]

[1]Kmbb college of engg Bhubaneswar,Odisha

[2]Kmbb college of engg Bhubaneswar,Odisha

[3]Kmbb college of engg Bhubaneswar,Odisha

Abstract

Breast cancer is the leading cause of cancer death among women. Screening mammography is the only method currently available for the reliable detection of early and potentially curable breast cancer. Research indicates that the mortality rate could decrease by 30% if women age 50 and older have regular mammograms. The detection rate can be increased 5-15% by providing the radiologist with results from a computer-aided diagnosis (CAD) system acting as a second opinion. However, among screening mammograms routinely interpreted by radiologists, very few (approximately 0.5%) cases actually have breast cancer. It would be beneficial if an accurate CAD system existed to identify normal mammograms and thus allowing the radiologist to focus on suspicious cases. This strategy could reduce the radiologist's workload and improve screening performance. Image mining is concerned with knowledge discovery in image databases. Since mammography is considered as the most effective means for breast cancer diagnosis, this paper introduces multi dimensional genetic association rule mining for classification of mammograms. The image Data mining approach has four major steps: Preprocessing, Feature Extraction, Preparation of Transactional database and multi dimensional genetic association rule mining. The purpose of our experiments is to explore the feasibility of data mining approach.. Results will show that there is promise in image mining based on multi dimensional genetic association rule mining. It is well known that data mining techniques are more suitable to larger databases than the one used for these preliminary tests. Computer-aided method using association rule could assist medical staff and improve the accuracy of mammogram detection. In particular, a Computer aided method based on association rules becomes more accurate with a larger dataset .Experimental results show that this new method can quickly and effectively mine potential association rules.

Keywords: Mammogram, Gray Level Co-occurrence Matrix features, Histogram Intensity, Classification, Genetic Algorithm; Association rule mining,

1. INTRODUCTION

Breast Cancer is one of the most common cancers, leading to cause of death among women, especially in developed countries. There is no primary prevention since cause is still not understood. So, early detection of the stage of cancer allows treatment which could lead to high survival rate. Mammography is currently the most effective imaging modality for breast cancer screening. However, 10-30% of breast cancers are missed at mammography [1]. Mining information and knowledge from large database has been recognized by many researchers as a key research topic in database system and machine learning Researches that use data mining approach in image learning can be found in [2,3].

Data mining of medical images is used to collect effective models, relations, rules, abnormalities and patterns from large volume of data. This procedure can accelerate the diagnosis process and decision-making. Different methods of data mining have been used to detect and classify anomalies in mammogram images such as wavelets [4,5], statistical methods and most of them used feature extracted using image processing techniques [6].Some other methods are based on fuzzy theory [7,8] and neural networks [9]. In this paper we have used classification method called a novel Multidimensional Genetic Association Rule Miner (MGARM) is proposed for rule construction. The result shows that the proposed rule-based approach reaches the classification accuracy over 96% and also demonstrates the use and effectiveness of association rule mining in image classification [10-12].

Classification process typically involves two phases: training phase and testing phase. In training phase the properties of typical image features are isolated and based on this training class is created .In the subsequent testing phase , these feature space partitions are used to classify the image. We have used supervised genetic association rule method by extracting low level image features for classification. The merits of this method are effective feature extraction, selection and efficient classification. The rest of the paper is organized as follows. Section 2 presents the pre-processing and section 3 presents the feature extraction phase. Section 4 discusses the proposed method of Feature selection and classification. In section 5 the results are discussed and conclusion is presented in section 6.

2. PRE-PROCESSING

The mammogram image for this study is taken from Mammography Image Analysis Society (MIAS)†, which is an UK research group organization related to the Breast cancer investigation [13]. As mammograms are difficult to interpret, pre-processing is necessary to improve the quality of image and make the feature extraction phase as an easier and reliable one. The calcification cluster/tumour is surrounded by breast tissue that masks the calcifications preventing accurate detection and shown in Figure.1. .A pre-processing; usually noise-reducing step is applied to improve image and calcification contrast figure 2. In this work [14] efficient filter referred to as the low pass filter was applied to the image that maintained calcifications while suppressing unimportant image features.

Figure 2 shows representative output image of the filter for a image cluster in Figure 1. By comparing the two images, we observe background mammography structures are removed while calcifications are preserved. This simplifies the further tumour detection step.

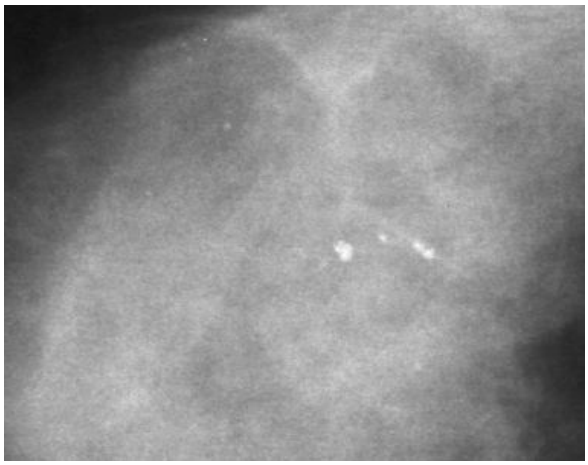

Figure 1 ROI of a Benign

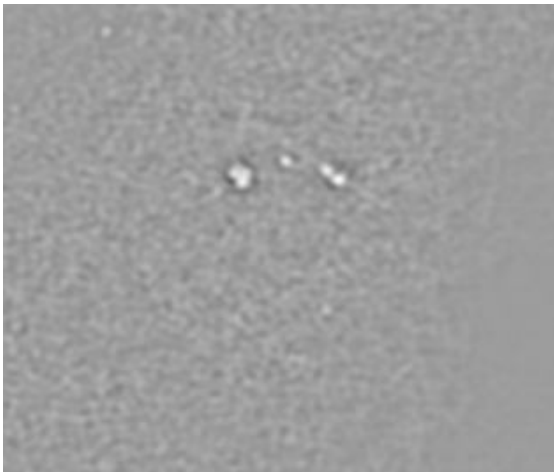

Figure 2 ROI after Pre-processing Operation

2.1 Histogram Equalization

Histogram equalization is a method in image processing of contrast adjustment using the image's histogram [15]. Through this adjustment, the intensities can be better distributed on the histogram. This allows for areas of lower local contrast to get better contrast. Histogram equalization accomplishes this by efficiently spreading out the most frequent intensity values. The method is useful in images with backgrounds and foregrounds that are both bright or both dark. In particular, the method can lead to better views of bone structure in x-ray images, and to better detail in photographs that are over or under-exposed. In mammogram images Histogram equalization is used to make contrast adjustment so that the image abnormalities will be better visible.

† peipa.essex.ac.uk/info/mias.html

3. FEATURE EXTRACTION

Features, characteristics of the objects of interest, if selected carefully are representative of the maximum relevant information that the image has to offer for a complete characterization a lesion [16, 17]. Feature extraction methodologies analyse objects and images to extract the most prominent features that are representative of the various classes of objects. Features are used as inputs to classifiers that assign them to the class that they represent.

In this Work intensity histogram features and Gray Level Co-Occurrence Matrix (GLCM) features are extracted.

3.1 Intensity Histogram Features

Intensity Histogram analysis has been extensively researched in the initial stages of development of this algorithm [18]. Prior studies have yielded the intensity histogram features like mean, variance, entropy etc. These are summarized in Table I Mean values characterize individual calcifications; Standard Deviations (SD) characterize the cluster. Table 2 summarizes the values for those features.

Table 1: Intensity histogram features

Feature Number assigned	Feature
1.	Mean
2.	Variance
3.	Skewness
4.	Kurtosis
5.	Entropy
6.	Energy

In this paper, the value obtained from our work for different type of image is given as follows:

Table 2: intensity histogram features and their values

Image Type	Features					
	Mean	Variance	Skewness	Kurtosis	Entropy	Energy
normal	7.2534	1.6909	-1.4745	7.8097	0.2504	1.5152
malignant	6.8175	4.0981	-1.3672	4.7321	0.1904	1.5555
benign	5.6279	3.1830	-1.4769	4.9638	0.2682	1.5690

3.2 GLCM Features

It is a statistical method that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix [19, 20, 21, 22]. By default, the spatial relationship is defined as the pixel of interest and the pixel to its immediate right (horizontally adjacent), but you can specify other spatial relationships between the two pixels. Each element (I, J) in the resultant GLCM is simply the sum of the number of times that the pixel with value I occurred in the specified spatial relationship to a pixel with value J in the input image.

The Following GLCM features were extracted in our research work:

Autocorrelation, Contrast, Correlation, Cluster Prominence, Cluster Shade, Dissimilarity Energy, Entropy, Homogeneity, Maximum probability, Sum of squares, Sum average, Sum variance, Sum entropy, Difference variance, Difference entropy, information measure of correlation1, information measure of correlation2, Inverse difference normalized. Information difference normalized. The value obtained for the above features from our work for a typical image is given in the following table 3.

Table3: GLCM Features and values Extracted from Mammogram IMAGE (Malignant)

Feature No	Feature Name	Feature Values
1	Autocorrelation	44.1530
2	Contrast	1.8927
3	Correlation	0.1592
4	Cluster Prominence	37.6933
5	Cluster Shade	4.2662
6	Dissimilarity	0.8877
7	Energy	0.1033
8	Entropy	2.6098
9	Homogeneity	0.6645
10	Maximum probability	0.6411
11	Sum of squares	0.1973,
12	Sum average	44.9329
13	Sum variance	13.2626
14	Sum entropy	133.5676
15	Difference variance	1.8188
16	Difference entropy	1.8927
17	Information measure of correlation1	1.2145
18	Information measure of correlation2	-0.0322
19	Inverse difference normalized	0.2863
20	Information difference normalized	0.9107

4. CLASSIFICATION

1. **Association rule mining:** Association rule mining is one of the important tasks of data mining intended towards decision support. Basically it is the process of finding some relations among the attributes of a huge database. Such relationships will help in taking some decisions. The process of extracting these relationships is termed as association rule mining. A number of algorithms have been developed for searching these rules [23, 24]. In this work, we have used the measures like information gain and interestingness [25, 26], used for constructing and evaluating a rule.

2. **Multi-Dimensional Genetic Association Rule Miner:** The GLCM features extracted for the digital mammograms are discretized using WEKA [27], an open source tool freely downloadable and available from http://www.cs.waikato.ac.nz/ml/weka and the discretized values are stored in database, in which, each columns represents one feature (attribute) and the last column represents the class attribute and tuples are used to represent images. A novel genetic algorithmic approach named Multidimensional Genetic Association Rule Miner (MGARM) is proposed for constructing rules for classification of MCs. Here, the multi-dimensional means, for each class in the database, a separate thread of GA is applied to construct the rule. Finally the best rule in each thread is combined to form the rule set.

The detail of the proposed algorithm is explained in the following sequence.

The genes are the basic elements of GA. Here the attributes are considered as genes. The sequence of genes is known as chromosome, represents one rule. A collection of 20 chromosomes generated for each population. The MCs are going to be classified into three types: normal, benign and malign. So, in this work, the GA has 3 dimensions, means 3 set of populations. The chromosomes

are encoded with numbers, every 2 digits are reserved for one attribute and the final attribute is mean for the class attribute. In this way the rule can be easily decoded. Each attribute can have set of possible values. The encoding represents one of the possible values with its index. If that gene 00 means that the corresponding attribute is not included in the rule. Consider the following chromosome:

00 11 05 00 01

Which represents a rule, in that, the first and fourth attributes are not considered, the second and third attributes are included with their 11_{th} and 5_{th} index values. For rule construction, the items are selected at random, and the information gain is calculated for the attribute, if it is greater than the threshold (0.5) then the item will be added to the rule, ignored otherwise. For each rule the interestingness measure is calculated as fitness value. And the rule having highest interestingness is stored as global best rule for each population. Then the genetic operators are applied to generate a new set of population as given:

Reproduction (selection)

The selection process selects chromosomes from the mating pool directed by the survival of the fittest concept of natural genetic systems. In the proportional selection strategy adopted in this article, a chromosome is assigned a number of copies, which is proportional to its fitness in the population that goes into the mating pool for further genetic operations. Roulette wheel selection is one common technique implements the selection strategy.

Crossover

Crossover is a probabilistic process that exchanges information between two parent chromosomes for generating two child chromosomes. In this paper, single point crossover with a fixed crossover probability of p_c=0.6 is used. For chromosomes of length *l*, a random integer, called the crossover point, is generated in the range [1, *l*-1]. The portions of the chromosomes lying to the right of the crossover point are exchanged to produce two offspring.

Mutation

Each chromosome undergoes mutation with a fixed probability p_m=0.03. For binary representation of chromosomes, a bit position is mutated by simply flipping its value. Since we are considering real numbers, a random position is chosen in the chromosome and replace by a random number between 0-9.

From the new set of populations, the best rules are extracted known as locally best. The global and local best rules are compared. If local rules are better than the next iteration is continued with the new populations and the local rules are saved as global best. Otherwise, the next iteration is performed with the old populations. The global best rules are pruned to check whether the quality is improving or not. In rule pruning, the attributes are temporarily removed one by one at random, if the interestingness measure improves than the attributes are removed permanently. The following algorithm describes our proposed method.

MGARM Algorithm

1. Load the training samples of GLCM features.
2. Construct 20 chromosomes (rules) based on information gain for M populations, one population for each dimension (class).
3. Calculate the interestingness of each chromosome; assign them as fitness value.
4. For each population, Store the chromosomes having better interestingness as global best rule.
5. Apply the Genetic operators such as reproduction, crossover and mutation to construct the new population.
6. For the new populations at each dimension, calculate the interestingness of each chromosome.
7. Choose the locally best rule.
8. Compare the global and local best rules.
9. If local rules are better than the global continue the next iteration with the new population and hold the local best rule as global best. Otherwise, continue with the old population.

And perform rule pruning.

Repeat from Step 5 for n number of iterations.

Decode the global best chromosomes to form the association rules for each class.

5. EXPERIMENTAL RESULTS

The digital mammograms used in our experiments were taken from the Mammographic Image Analysis Society (MIAS). The database consists of 322 images, which belong to three categories: normal, benign and malign (ftp://peipa.essex.ac.uk). There are 208 normal images, 63 benign and 51 malign, which are considered abnormal.

The proposed method is evaluated based on ten-fold cross validation method. The following table presents the rule accuracy of the proposed classification system compared with other association rule based system proposed in [25, 26]. The results for the ten splits of the mammogram database are given in Table 4.

Table 4: Classification accuracy for the ten splits with MGARM

Splits	Classification Accuracy
1	98.65
2	95.89
3	97.76
4	93.12
5	92.23
6	98.66

7	94.41
8	98.54
9	98.05
10	95.89
Average	**96.32**

In this paper we used multi dimensional genetic association rule mining using image contents for the classification of mammograms. The average accuracy is 96.32 %. We have employed the freely available Machine Learning package, WEKA [27]. Out of 322 images in the dataset, 230 were used for training and the remaining 92 for testing purposes and the result is shown in Table 5.

Table 5: Results obtained by proposed method

Normal	100%
Malignant	88. 23%
Benign	97.11%

The confusion matrix has been obtained from the testing part .In this case for example out of 51 actual malignant images 06 images was classified as normal. In case of benign all images are correctly classified and in case of normal images 6 images are classified as malignant. The confusion matrix is given in Table 6.

Table 6: Confusion matrix

Actual	**Predicted class**		
	Benign	**Malignant**	**Normal**
Benign	63	0	0
Malignant	51	45	06
Normal	208	6	202

6. CONCLUSION

Automated breast cancer detection has been studied for more than two decades Mammography is one of the best methods in breast cancer detection, but in some cases radiologists face difficulty in directing the tumors. We have described a comprehensive of methods in a uniform terminology, to define general properties and requirements of local techniques, to enable the readers to select the efficient method that is optimal for the specific application in detection of micro calcifications in mammogram images. Classification of Microcalcification Clusters (MCs) is one the key to find the early sign of breast cancer. In this paper, we have proposed a novel association rule based system for classification of Microcalcification Clusters (MCs). Initially the MCs are segmented from the mammograms and the statistical GLCM features are extracted. The proposed approach Multi Multidimensional Genetic Association Rule Miner (MGARM) is applied to construct the association rule to classify the images into three classes: normal, benign and malign. The result shows that MGARM outperforms than the existing. In future, an efficient algorithm can be used to select the relevant features and the rules can be generated to improve the accuracy.

References

[1] Majid AS, de Paredes ES, Doherty RD, Sharma N Salvador X. “Missed breast carcinoma: pitfalls and Pearls”. Radiographics, pp.881-895, 2003

[2] Osmar R. Zaïane,M-L. Antonie, A. Coman “Mammography Classification by Association Rule based Classifier,” MDM/KDD2002 International Workshop on Multimedia Data Mining ACM SIGKDD, pp.62-69,2002

[3] Xie Xuanyang, Gong Yuchang, Wan Shouhong, Li Xi Computer Aided Detection of SARS Based on ,” Radiographs Data Mining “, Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, pp7459 – 7462, 2005

[4] C.Chen and G.Lee, “Image segmentation using multitiresolution wavelet analysis and Expectation Maximum(EM) algorithm for mammography” , International Journal of Imaging System and Technology, 8(5): pp491-504,1997.

[5] T.Wang and N.Karayaiannis, “Detection of microcalcification in digital mammograms using wavelets”, IEEE Trans. Medical Imaging, 17(4):498-509, 1998

[6] Jelena Bozek, Mario Mustra, Kresimir Delac, and Mislav Grgic “A Survey of Image Processing Algorithms in Digital mammography”Grgic et al. (Eds.): Rec. Advan. in Mult. Sig. Process. and Commun., SCI 231, pp. 631–657,2009

[7] Shuyan Wang, Mingquan Zhou and Guohua Geng, “Application of Fuzzy Cluster analysis for Medical Image Data Mining” Proceedings of the IEEE International Conference on Mechatronics & Automation Niagara Falls, Canada,pp. 36 – 41, 2005.

[8] Jensen, Qiang Shen, “Semantics Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches”, IEEE Transactions on Knowledge and Data Engineering, pp. 1457-1471, 2004

[9] I.Christiyanni et al ., “Fast detection of masses in computer aided mammography”, IEEE Signal processing Magazine, pp:54- 64,2000.

[10] K. Thangavel , A. Kaja Mohideen "Classification of Microcalcifications Using Multi-Dimensional Genetic Association Rule Miner" International Journal of Recent Trends in Engineering, Vol 2, No. 2, pp. 233 – 235, 2009

[11] R.Agrawal, T. Imielinski, and A.Swami. Mining association rules between sets of items in large databases. In the Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (ACM SIGMOD '93), Washington, USA, May 1993

[12] Alex A. Freitas, "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery" Postgraduate Program in Computer Science, Pontificia Universidade Catolica do Parana Rua Imaculada Conceicao, 1155. Curitiba - PR. 80215-901. Brazil.

[13] Etta D. Pisano, Elodia B. Cole Bradley, M. Hemminger, Martin J. Yaffe, Stephen R. Aylward, Andrew D. A. Maidment, R. Eugene Johnston, Mark B. Williams,Loren T. Niklason, Emily F. Conant, Laurie L. Fajardo,Daniel B. Kopans, Marylee E. Brown • Stephen M. Pizer "Image Processing Algorithms for Digital Mammography: A Pictorial Essay" journal of Radio Graphics Volume 20,Number 5,sept.2000

[14] Pisano ED, Gatsonis C, Hendrick E et al. "Diagnostic performance of digital versus film mammography for breast-cancer screening". NEngl J Med 2005; 353(17):1773-83.

[15] Wanga X, Wong BS, Guan TC. 'Image enhancement for radiography inspection". International Conference on Experimental Mechanics. 2004: 462-8.

[16] D.Brazokovic and M.Nescovic, "Mammogram screening using multisolution based image segmentation", International journal of pattern recognition and Artificial Intelligence, 7(6): pp.1437-1460, 1993

[17] Dougherty J, Kohavi R, Sahami M. "Supervised and unsupervised discretization of continuous features". In: Proceedings of the 12th international conference on machine learning.San Francisco:Morgan Kaufmann; pp 194–202, 1995.

[18] Yvan Saeys, Thomas Abeel, Yves Van de Peer "Towards robust feature selection techniques", www.bioinformatics.psb.ugent

[19] Gianluca Bontempi, Benjamin Haibe-Kains "Feature selection methods for mining bioinformatics data", http://www.ulb.ac.be/di/mlg

[20] Li Liu, Jian Wang and Kai He "Breast density classification using histogram moments of multiple resolution mammograms" Biomedical Engineering and Informatics (BMEI), 3rd International Conference, IEEE explore pp.146–149, DOI: November 2010, 10.1109/ BMEI.2010 .5639662,

[21] Li Ke,Nannan Mu,Yan Kang Mass computer-aided diagnosis method in mammogram based on texture features, Biomedical Engineering and Informatics (BMEI), 3rd International Conference, IEEE Explore, pp.146 – 149, November 2010, DOI: 10.1109/ BMEI.2010.5639662.

[22] Azlindawaty Mohd Khuzi, R. Besar and W. M. D. Wan Zaki "Texture Features Selection for Masses Detection In Digital Mammogram" 4th Kuala Lumpur International Conference on Biomedical Engineering 2008 IFMBE Proceedings, 2008, Volume 21, Part 3, Part 8, 629-632, DOI: 10.1007/978-3-540-69139-6_157

[23] J Hipp, U Güntzer, and G Nakhaeizadeh, "Algorithms for association rule mining—a general survey and comparison", vol. 2, no. 1, 2000.

[24] Jiawei Han and Micheline Kamber, "Data Mining, Concepts and Techniques". Morgan Kaufmann, 2001.

[25] ML Antonie, OR. Zaiane, and A Coman, "Application of data mining techniques for medical image classification". In Proc. Of Second Intl. Workshop on Multimedia Data Mining (MDM/KDD'2001) in conjunction with Seventh ACM SIGKDD, pp 94–101, San Francisco, USA, 2001.

[26] Deepa S. Deshpande "ASSOCIATION RULE MINING BASED ON IMAGE CONTENT" International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 143-146

[27] Holmes, G., Donkin, A., Witten, I.H.: WEKA: a machine learning workbench. In: Proceedings Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia, pp. 357-361, 1994.

AUTHOR

Aswini Kumar Mohanty received the B.E. and M.TECH. Degree in Computer Engineering from Marathawada University and Kalinga University in 1991 and 2005 respectively. He completed Phd in Information Technology in 2013 from Vinoba Bhave University. He has more than 24 years experience in teaching and industry. His area of research is medical image processing and data mining. Presently he is working as principal in KMBB College of engineering, khurda, odisha.

Amalendu Bag received the MCA and M.TECH. Degree in Computer Science from NIT, Rourkela and Biju Patnaik University of Technology in 2003 and 2012 respectively. He has more than 12 years experience in teaching and industry. His area of research is medical image processing and soft computing. Presently he is working as Assistant Professor in Computer Science Deptt. in KMBB College of engineering, khurda, odisha.

Devitosh Acharya received the MCA and M.TECH. Degree in Computer Science from Madurai Kamaraj and KIT University in 2001 and 2006 respectively. He has more than 14 years experience in teaching and industry. His area of research is medical image processing and soft computing, Embedded System, Real time system. Presently he is working as Assistant Professor in Computer Science Deptt. in KMBB College of engineering, khurda, odisha.