Mass Detection and Classification using Machine Learning Techniques in Digital Mammograms

S Narasimha Murthy
PET Research Centre
PES College of Engg.
Mandya, Karnataka

Arun Kumar M N
PET Research Centre
PES College of Engg.
Mandya, Karnataka

H S Sheshadri
PET Research Centre
PES College of Engg.
Mandya, Karnataka

ABSTRACT
Breast cancer is one of the most dangerous carcinomas for middle-aged and older women in the world. Mammography is a detection tool that assists the radiologists in reading the mammograms. In this paper, new techniques are proposed to detect and classify the masses automatically. These techniques improve the detection and classification process. Classification of masses into benign or malignant is an issue as the number of instances belongs to benign class is significantly greater than the malignant classes. This imbalanced problem is well addressed in proposed method using different approaches. This classification method outperforms many other classification approaches.

Keywords
Classification, Masses, Imbalanced data sets, Digital Mammography

1. INTRODUCTION
Breast cancer [1-4] is the most frequently occurring cancer and one of the leading causes of death among women. Early diagnosis and subsequent treatment can significantly improve the chance of survival for patients with breast cancer. Effective method for the detection of early breast cancer is mammography. Due to the subtle nature and poor image quality mammograms are the most difficult radiological images to interpret by radiologists. Mammography images are often very poor in contrast and can show different features and patterns depending on breast anatomy and tissues density. Radiologists do not detect all breast cancers that are retrospectively detected on the mammograms. Mammogram screening acts as a "second reader" to the physician. Computer-Aided Detection (CADe) and Computer-Aided Diagnosis (CADx) are the two major processes involved in the development of techniques that perform analysis and interpretation of breast mammogram. Detection identifies potential abnormalities, such as microcalcification, masses, and architectural distortions. Diagnosis classifies a detected abnormality as benign or malignant. Pre-processing, segmentation, feature extraction, and classification are a series of tasks done before CADe algorithms can identify suspicious regions in a mammogram.

Breast cancer [5] is one of the most dangerous carcinomas for middle-aged and older women in the world. As said earlier mammography is the most effective and reliable detection tool of breast cancer. Many computer aided diagnosis (CAD) methods have been developed for analyzing and interpreting the mammogram. In the mammograms [6-8], masses are the most important symptoms of abnormality. Accurate diagnosis of masses is very difficult because of their appearances. In order to accurately detect the abnormality hidden inside the breast tissue, a method [9] which incorporates mainly a DOG (difference of Gaussian) approach is used to locate the masses. To remove the artefacts they used a filtering process based on area, contrast and roundness. Authors in paper [10] used fractal dimension and a multi-resolution Markov random field (MRF) to detect the suspicious regions and to find the masses. Authors [11] used Radon transform to enhance the digitized mammogram image, and a series of linear radial spiculation filters (RSF) are to detect the spiculated masses. With these CAD methods, a comparatively high detection precision could be achieved, if the masses always had ordinary characteristics and backgrounds. If the features of the masses were special or their backgrounds were different, the diagnosis precision would be reduced acutely, and the false positive (FP) rate can hardly be suppressed, because none of these conventional had adequate automatic adjustment ability. There are many methods developed for the imbalanced data classification and very little efforts have been carried out for the classification of masses using imbalanced data classification approaches.

There are some approaches for the imbalanced data classification. Information granulation based data mining is proposed in [12]. Two balanced datasets and unbalanced datasets from UCI machine learning repository are used for experimental analysis. They compared the effectiveness of the proposed method with NN. Authors reported that their method outperforms the NN. They concluded their work with the remark that to reduce the data dimension more sophisticated tool can be effectively used. SVM based active learning (AL) selection strategy is developed in [13]. 8 datasets of Reuters, 5 datasets from Citeseer, 3 datasets from UCI, USPS, MNIST-8 are used for experimental work. SMOTE and DC are the other methods used in the comparative analysis and AL outperforms in 12 out of 18 cases and there is no risk of losing the information. Future research work can be focused on medical images with intractable geometric complexity in data classification.

Authors in [14] proposed a combination of supervised learning and unsupervised clustering to handle the imbalanced data set. Proposed methods are applied on Liver, Hepatitis, Pima diabetes, and Wisconsin datasets. Comparative analysis with RPROP is carried out and the proposed method has higher values of G-mean and F-measure than RPROP but similar classification rates for both the methods. The limitation in their work is that they have not studied the efficiency of the method on imbalanced data sets of very highly less imbalanced ratio < 0.1. Four combination of

SMOTE and complementary neural network (CMTNN) to handle problem of classifying imbalanced data is proposed in [15]. Authors used ANN, k-NN, and SVM for the classification of imbalanced data. Datasets like Pima Indian diabetes data, German credit data, Haberman's survival data, SPECT Heart data are used for experimental analysis. Performance comparison is done with other methods like ENN, Tomek links, and SMOTE. It is observed that for ANN classifier, four combination of SMOTE and CMTNN outperforms others. For SVM classifier Tomeklinks and one combination of (SMOTE+CMTNN) outperform other. For a k-NN classifier, two combination of SMOTE + CMTNN outperforms the other methods. Cost sensitive boosting algorithm (AdaC2.M1) is proposed [16] by Yanmin Sun et.al. Car data, New thyroid data, and Nursery data from UCI machine learning database are used for the experimental analysis. AdaC2.M1 is compared with Adaboost M1 and it is found that AdaC2M1 outperforms Adaboost.M1.

In this paper, a new mass detection and classification algorithm is proposed. It uses morphological operation to detect the masses and imbalanced classification methods to classify the masses and method could get high detection precision and low FP rate. Paper is organized as follows. Section 2 explains the materials used and the Pre-processing of mammograms. Section 3 explains the detection and classification of masses. Experimental Results are discussed in Section 4. Conclusions are drawn in section 5.

2. MATERIALS AND PREPROCESSING

In this experiment, totally 20 cases were used to evaluate the detection and classification effect of the masses. All of these mammograms were taken from the MIAS. Before the formal detection process of the masses was carried out, some preprocessing steps are carried out and breast regions are extracted.

Preprocessing:-

(i). Breast Border Detection. Proposed method used an existing similar technique [5] for breast region segmentation using morphological and filtering techniques. The steps followed to detect the breast border involves: - Removal of noise, morphological operation, Edge detection , filtering, breast border extraction.

Removal of Noise

Nonlinear filter is used to remove the impulsive noise from an image. In the proposed work a median filter is used in which output pixel contains the median value in the 5X5 neighborhood around the corresponding pixel in the input image.

Morphological Operation

The original mammogram is opened by using a suitable structuring element. Image is thresholded with a suitable threshold value, which is experimentally obtained. Morphological operations are applied to smooth the image.

Edge Detection and Filtering Techniques

This step uses the Canny edge detection and then dithering and 2-D order statistic filtering.

Multidimensional image filtering

This step removes the noises using a multidimensional image filtering. A Gaussian low pass filter filters the image. Finally the image is converted to binary image and erosion is carried out.

Finally binary image perimeter pixels are found. This perimeter is the boundary of the breast image. A pixel is the part of the perimeter if it is nonzero and it is connected to at least one zero-valued pixel. The connectivity used is 8.

(ii) Locate the region containing the pectoral muscle

Pectoral muscle detection is a challenging task in the breast segmentation process. An existing algorithm [5] is used for pectoral muscle segmentation. Technique for segmenting pectoral muscle presented uses wavelet decomposition, and edge detection using the Sobel. The ROI containing pectoral muscle is determined in two steps. A rectangle which encloses the pectoral muscle is determined first and then a refinement/reduction to this rectangle is performed to reduce the processing time.

3. MASS DETECTION AND CLASSIFICATION

This section explains the approaches used for mass detection and classification.

3.1 Mass detection

This step removes the background without reducing the masses by a top hat filtering, compute optimal thresholds for segmenting the image. The steps are

1. Filter the image by a top hat filter.

2. For all the grey levels Gs compute the Image Entropy (IE). Store IE and the Gs used in the arrays X and Y respectively.

3. Rank the array X (i.e., obtain the descending order of X) and then rank the corresponding Y array. Array Y stores the optimal thresholds (grey levels) in the descending order.

To detect greater possible quantity of masses the top 5 optimal thresholds are selected (experimentally obtained) and used to segment the filtered image. When each of these thresholds is applied the newly obtained pixels are added and the repeated pixels are deleted. With the available truth information in MIAS database and with the support of 3 radiologists these segmented area were classified as benign and malignant masses. From these detected segments a high percentages indicate a class imbalance problem and this issue is addressed by using an improved classifier that introduces balanced learning for the accurate classification of masses.

3.2 Mass Classification

To accurately classify the masses a set of features are extracted and the dataset is pre-processed with SafelevelSMOTE, SMOTE and Borderline SMOTE. Three classifiers, SVM, Bayesian Networks, and kNN are used for the classification of the pre-processed dataset.

Safe-Level-Synthetic Minority Oversampling TEchnique assigns each positive instance its safe level before generating synthetic instances. Each synthetic instance is positioned closer to the largest safe level so all synthetic instances are generated only in safe regions. The instance is nearly noise if the safe level of an instance is close to 0. Safe-Level-SMOTE algorithm is shown in figure 1.

Description of variables used in algorithm

p is an instance in the set of all original positive instances A.

n is a selected nearest neighbours of p.

s is a synthetic instance.

bln is safe level of p

xy is safe level of n

sl_ratio is safe level ratio.

NUMAT is the number of attributes.

DIFFERENCE is the difference between the values of n and p at the same attribute id.

gap is a random fraction of DIFFERENCE.

|A| is the number of all positive instances in A

A' is a set of all synthetic instances returned when the algorithm terminates

Input: A set of all original positive instances A

Output: A set of all synthetic positive instances A'

1. A' = ∅

2. for each positive instance p in A {

3. compute k nearest neighbors for p in A and

randomly select one from the k nearest neighbors, call it n

4. bln = the number of positive stances in k nearest neighbors for p in A

5. xy = the number of positive stances in k nearest neighbors for n in A

6. if (xy ≠ 0) { ; sl is safe level.

7. sl_ratio = bln / xy ; sl_ratio is safe level ratio.

8. }

9. else {

10. sl_ratio = ∞

11. }

12. if (sl_ratio = ∞ AND bln = 0) { ; the 1st case

13. does not generate positive synthetic instance

14. }

15. else {

16. for (atti = 1 to NUMAT) { ; NUMAT is the number of attributes.

17. if (sl_ratio = ∞ AND bln ≠ 0) { ; the 2nd case

18. gap = 0

19. }

20. else if (sl_ratio = 1) { ; the 3rd case

21. generate a random number between 0 and 1, call it gap

22. }

23. else if (sl_ratio > 1) { ; the 4th case

24. generate a random number between 0 and 1/sl_ratio, call it gap

25. }

26. else if (sl_ratio < 1) { ; the 5th case

27. generate a random number between 1-sl_ratio and 1, call it gap

28. }

29. DIFFERENCE = n[atti] - p[atti]

30. s[atti] = p[atti] + gap * DIFFERENCE

31. }

32. A' = A' ∪ {s}

33. }

34. }

35. return A'.

Fig 1: Algorithm: Safe-Level-SMOTE

Imbalanced data sets are also preprocessed using SMOTE and Borderline SMOTE algorithms before the classifiers are trained. Borderline-SMOTE are different from many over-sampling methods in which all the minority examples or a random subset of the minority class are over-sampled. It is based on SMOTE (Synthetic Minority Over-sampling Technique).

4. EXPERIMENTAL RESULTS

The main objective of the study is to evaluate the performance of the different classifiers. For data sampling proposed method used SMOTE, Borderline SMOTE, and Safelevel SMOTE. In order to demonstrate that the proposed sampling techniques can assist classification of imbalanced data, several classification algorithms are used. They are SVM, Bayesian Networks, and k-NN. The results of classification are evaluated and compared in terms of performance. Table 1 shows the comparison results.

Table 1. Comparison results

Sampling methods	Classifier	Sensitivity	Specificity
Borderline SMOTE	Bayesian	71.91	82.0
	k-NN	75.06	83.2
	SVM	81.3	80.76
Safelevel SMOTE	Bayesian	72.56	83.31
	k-NN	76.45	84.23
	SVM	84.34	83.31
SMOTE	k-NN	82.3	85.91
	ANN	79.4	86.01
	SVM	**86.4**	81.5

5. REFERENCES

[1] S. Oporto-D´ıaz, R. R. Hernandez-Cisneros and H. Terashima-Mar´ın, ―Detection of microcalcification clusters in mammograms using a difference of optimized Gaussian filters‖, in Proceedings of the Second International Conference on Image Analysis and Recognition, ICIAR 2005, Toronto, ON, Canada, pp. 998–1005, 2005.

[2] Karssemeijer, N and Hendrikis, L. (1997). Computer assisted reading of mammograms Eur. Radiol. (7), 743-748

[3] Kim, J, K and Park H. W. (1999). Statistical textural features for detection of microcalcifications in digitized mammograms. IEEE Transactions on Medical Imaging (18), 231-238

[4] Mushlin, R and Shapiro, K, D.(1998). Estimating the Accuracy of screening mammography: A meta analysis. Journal of Preventive Medicine vol.14 (2)143-153

[5] Arun kumar M.N and H.S. Sheshadri, ―Breast contour extraction and pectoral muscle segmentation in digital mammograms‖, International Journal of Computer Science and Information Security, Vol 9, No.2, February 2011.

[6] Rolando R. Hern´andez-Cisneros and Hugo Terashima, ―Evolutionary Neural Networks Applied To The Classification Of Microcalcification Clusters In Digital Mammograms‖, 2006 IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall

[7] Arun kumar M.N & Dr. H.S Sheshadri, "Building Accurate Classifier for the Classification of Microcalcification", published in International Journal of Computer Science and Information Technologies, Vol. 3 (6), pp.5346-5350, October 2012.

[8] Arun kumar M.N & Dr. H.S Sheshadri, "On the Classification of Imbalanced Datasets", published in International Journal of Computer Applications (IJCA), DOI: 10.5120/6280-8449, Volume 44– No.8, April 2012.

[9] W.E. Polakowski, D.A. Cournoyer, and S.K. Rogers, et al, "Computer aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency," IEEE Transactions on Medical Imaging, vol. 16, pp. 811-819, 1997.

[10] L. Zheng, and A.K Chan, "An artificial intelligent algorithm for tumor detection in screening mammogram," IEEE Transactions on Medical Imaging, vol. 20, pp. 559-567, 2001.

[11] M.P. Sampat, and A.C. Bovik, "Detection of spiculated lesions in mammograms," 25th IEEE Annual International Conference of the EMBS, Cancun, Mexico, pp. 810-813, 2003

[12] Mu-Chen Chen a, Long-Sheng Chen, Chun-Chin Hsu, Wei-Rong Zeng , ―An information granulation based data mining approach for classifying imbalanced data ―, Elsevier, Information Sciences 178 (2008) 3214–3227.

[13] Jian Huang, L´eon Bottou, C. Lee Giles, ―Learning on the border: Active learning in imbalanced data classification ―, CIKM'07, November 6–8, 2007, Lisboa, Portugal. ACM 978-1-59593-803-9/07/0011

[14] Son Lam Phung, Abdesselam Bouzerdoum, Giang Hoang Nguyen, ―Learning pattern classification tasks with imbalanced data sets ―, http://ro.uow.edu.au

[15] Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung, ―Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm‖, http://researchrepository.murdoch.edu.au

[16] Yanmin Sun, Mohamed S Kamel, and Yang Wang, ―Boosting for learning multiple classes with imbalanced class distribution ―, Proceedings of the Sixth International Conference on Data Mining (ICDM'06), 0-7695-2701-9/06 © 2006 IEEE.