Classification of Mammograms using Gray-level Co-occurrence Matrix and Support Vector Machine Classifier

P.Samyuktha,Vasavi College of engineering,CSE dept.

D.Sriharsha, IDD, Comp. Sc. & Engg., IIT (BHU), Varanasi (INDIA)

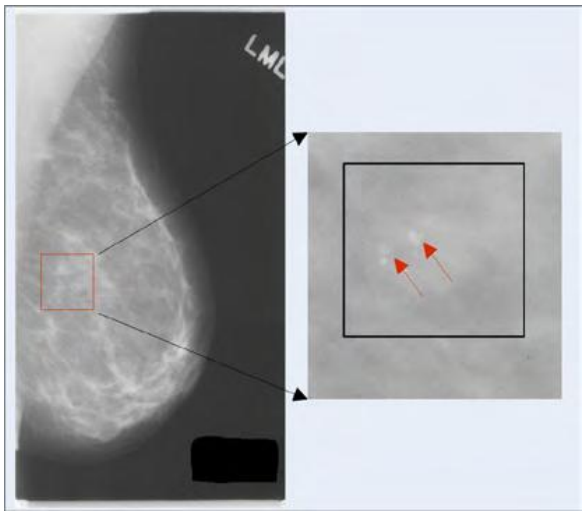
Abstract

Computer Aided Diagnosis (CAD) tools are used as second opinion by many radiologists in classifying various type of breast cancers. It already proved its success not only in reducing human error in reading the mammograms but also shows better and reliable classification into benign and malignant abnormalities. Here, this is an attempt to use Support Vector Machine (SVM) classifier for classification of mammograms based on Gray-level Co-occurrence Matrix (GLCM) texture based features.

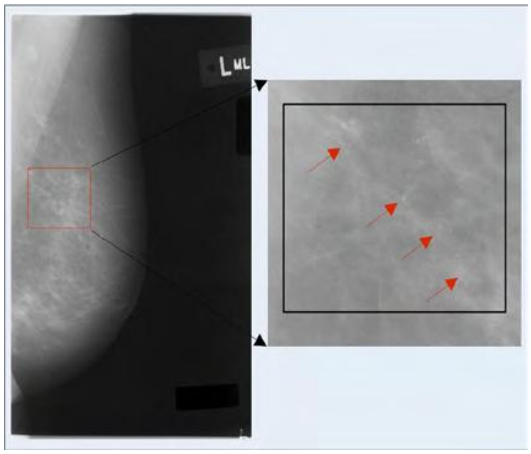
I. Introduction

In terms of cancer, breast cancer is the second killer after lung cancer. It is found mostly in women.A mammogram is a low dose x-ray picture of the breast. Mammograms can be used to check for breast cancer in women who have no signs or symptoms of the disease. This type of mammogram is called a screening mammogram. The x-ray images make it possible to detect tumors that cannot be felt. Screening mammograms can also find micro calcifications (tiny deposits of calcium) that sometimes indicate the presence of breast cancer. The women mortality rate due to the breast cancer can be reduced by early detection of breast cancer from screening mammograms. Mammograms can also be used to check for breast cancer after a lump or other sign or symptom of the disease has been found. This type of mammogram is called a diagnostic mammogram. Besides a lump, signs of breast cancer can include breast pain, thickening of the skin of the breast, nipple discharge, or a change in breast size or shape; however, these signs may also be signs of benign conditions. A diagnostic mammogram can also be used to evaluate changes found during a screening mammogram or to view breast tissue when it is difficult to obtain a screening mammogram because of special circumstances, such as the presence of breast implants.

The most common breast abnormalities that may indicate breast cancer include masses, calcifications, architectural distortion and bilateral symmetry. This dangerous illness is caused by lesion that can be classified into two categories, which are benign and malignant. Benign is harmless lesion which can be removed and unlikely to recur, while malignant is cancerous cell which highly potential to grow up and spread to other parts of the body. Microcalcification are also basically calcium deposits, but they are much smaller and much less common. Microcalcification tends to be the result of a genetic mutations somewhere in the breast tissue, but they can still be due to other conditions. The size, distribution, form, and density of microcalcifications are thought to give clues as to the potentially malignant nature of their origin. Most of breast cancer patient do not notice about its presence and died before they get proper medication. Thus, breast cancer detection on the early stage is necessary in order to reduce the number of deaths. So, many CAD systems are being proposed to detect this cancer in early stage itself. However, the exact classification into benign and malignant by these CAD tools had become difficult due to variations in tissue characteristics, such as shape, gray level, location, size, intensity, etc.Besides, the specificity of visual inspection is low with high sensitivity. In order to reduce the number of false-negative, diagnosis greater than 2% change of being malignant will be recommended to biopsy. Thus, an improvement in interpreting mammograms becomes an important issue.



BENIGN MICROCALCIFICATION



MALIGNANT MICROCALCIFICATION

Further, the breast abnormalities are defined with wide range of features and may be easily missed or misinterpreted by radiologists while examining large numbers of mammographic images (mammograms) provided in screening programs. Due to the limitations of human observers and its difficulty for radiologist to provide both accurate and uniform evaluation for the enormous number of mammograms generated in widespread screening, here in this paper the problem of design and development of a software CAD tool is addressed for the automation of the detection and diagnosis process. The CAD algorithms help reducing the number of false positives and they assist radiologists in deciding between follow up and biopsy and may be used as a second opinion. Hence, the design and development of an efficient automated CAD tool plays a major role. The steps involved in the design of a CAD tool for early breast cancer detection from mammograms include pre-processing, segmentation, feature extraction and selection, and classification.

Here, this is my attempt to classify these mammograms into normal or abnormal, by using the sequence of steps those are mentioned below.

II. Materials and Methods

2.1. Materials

Data set used in this study is digital mammograms taken from the published MIAS database (http://peipa.essex.ac.uk/ipa/pix/mias/). In this database, the original MIAS database are digitized at 50 micron pixel edge and has been reduced to 200 micron pixel edge and clipped or padded so that every image is 1024 X 1024 pixels. All images are held as 8-bit gray level scale images with 256 different gray levels (0-255) and physically in portable gray map (pgm) format. In total, MIAS database consists of 330 images which divided into 208 normal images, 68 benign images, and 54 malignant images. So, there are 64% of normal data, 20% of benign data, and 16% of malignant data. The data set is separated into training and testing with the comparison of 85:15. This can also be extended to data sets of much bigger size subject to availability, like Digital Database for Screening Mammography (DDSM) dataset etc.

2.2. Methods

This study is divided into three main processes, which are pre-processing, feature extraction, and classification. Classifier used for classification is Support Vector Machine (SVM)classifier. Kernel function svmtrain uses to map the training data into kernel space. The default kernel function used to train is linear(which uses dot product). In addition to linear, I also used rbf(Gaussian Radial Basis Function), and polynomial kernel (default uses order 3)

2.2.1. Pre-processing

The purpose of data pre-processing is to enhance the quality of data set by reducing irrelevant data that potentially interfere the training process. In this study, three irrelevant data are discarded from training set since the centroid of mass is not specified. They are 059.pgm, 212.pgm, and 214.pgm whose type is benign. And there are also multiple locations of calcifications in some of the mammograms in the dataset. Therefore, total data used in this study is 327 digital mammograms. Some process also performed in the processing step, which

are cropping on the Region of Interest (ROI) given along with images in dataset and resizing an image of a mammogram to be 128 x 128 pixels. The location of clusters of microcalcifications in a mammogram is mentioned in the data set as radius of the cluster and center point of the cluster. So images are cropped likewise and the mammograms with no clusters or normal ones are cropped at the center of the mammogram so that size of each mammogram be 128 x 128 pixels from 1024 x 1024 pixels.

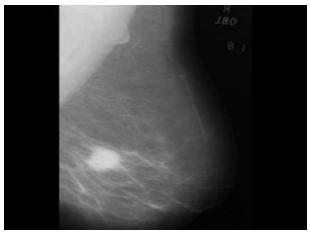


Image beforepre-processing

Image after pre-processing

2.2.2. Feature Extraction

Feature extraction is important in the process of getting the meaningful characteristic or information used in classification. Gray-level co-occurrence matrix (GLCM) is the technique to evaluate textures by considering the spatial relationship of the pixels. This method calculates the occurrence of pairs of pixels with specific values and in a specified spatial relationship in an image. The spatial relationship is between the pixel of interest and the pixel to its immediate right (horizontally adjacent). After that, it will produce the statistical method from the calculation matrix. The GLCM used in this experiment

calculate the occurrence of gray-level value *i*in specific spatial relationship to a pixel of *j* and then sum the number of *i*appears in the specific spatial relationship to pixel with value of *j* in the image .

In this experiment, feature extraction is useful to isolate either normal – abnormal lesion classification. Several prior studies show that the GLCM is an effective method for image texture analysis. The GLCM has been applied in several researches regarding image texture analysis and still become the significant aspect for the research. For this reason, GLCM method is used for digital mammogram texture extraction here.

The matrices are constructed at a distance of d = 1 and for direction of θ given as 0°, 45°, 90° and 135° in default. Then, four directions are used to extract the texture information for each masses and non-masses tiles area. The texture descriptors derived from GLCM are contrast, energy, homogeneity and correlation of gray level values. Haralick shows that there are 14 textural features. But, according to studies, 4 dominant features of GLCM based on t-test are Contrast, Correlation, Energy, and Homogeneity. Thus, this experiment will use only those 4 features.

Contrast -Measures the local variations in the gray-level co-occurrence matrix

Correlation -Measures the joint probability occurrence of the specified pixel pairs.

Energy -Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment.

Homogeneity - Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

Here, we use features of all the mammograms extracted from matrices that are constructed at a distance of d = 1, d = 3, d = 5, d = 7.

2.2.3. Validity

The accuracy rate of each classifier is evaluated using confusion matrix. For comparing the performance ofcancer detection of each classifier, specificity and sensitivity is used. Sensitivity and specificity are statistics used to measure the significance of a test related to the presence or absence of the disease. Equations below are used tocalculate these three parameters, respectively.

ACCURACY =

(TP+TN)*100 / (TP+TN+FP+FN)

SENSITIVITY=TP / (TP + FN)

SPECIFICITY = TN/ (FP+TN)

Where, TP is true positive, FN is false negative, TN is true negative, and FP is false positive. Based on the above equations it can be concluded that sensitivity indicates the number of disease that is correctly predictedby the positive test while specificity indicates the number of patients without disease who test negative. Thus, it is ameasure of test performance whose purpose is to distinguish between patients who do and do not suffer from thedisease.

2.2.4 Support Vector Machine Classifier

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering and is often used in industrial applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.

The diagrams below gives us the idea of working model of SVM classifier

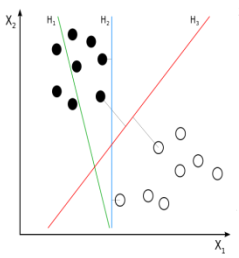


Figure1

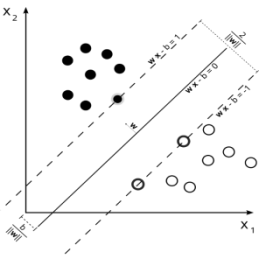


Figure 2

In figure 1, H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximum margin. Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a newdata point will be in. In the case of support vector machines, a data point is viewed as an n-dimensional vector (a list of n numbers), and we want to know whether we can separate such points with a (n-1)-dimensional hyperplane. This is called a linear classifier. There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum margin classifier or equivalently, the perceptron of optimal stability.

Figure 2 shows the working of linear SVM, in which (w.x – b = -1) and (w.x – b = 1). Any hyperplane can be written as the set of points x, satisfying (w.x – b = 0)

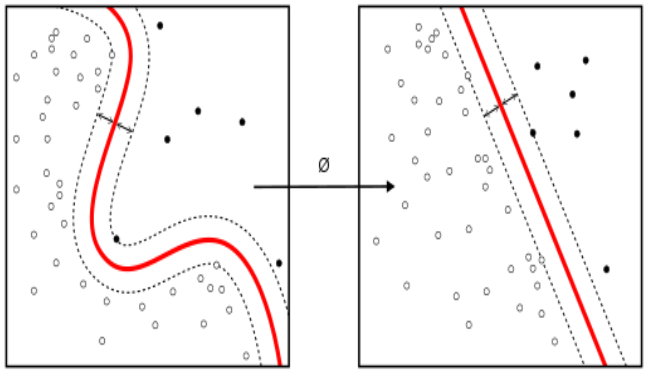


Figure 3

The original maximum-margin hyperplane algorithm proposed by Vapnik in 1963 constructed a linear classifier.However, in 1992, Bernhard E. Boser, Isabelle M. Guyon and VladimirNVapnik suggested a way to create nonlinear classifiers by applying the kernel trick(originally proposed by Aizerman) to maximum-margin hyperplanes. The above mentioned figure 3 shows us the working of Nonlinear SVM classifier. The resulting algorithm is formally similar, except that every dot product is replaced by a nonlinear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation may

be nonlinear and the transformed space high dimensional; although the classifier is a hyperplane in the transformed feature space, it may be nonlinear in the original input space.

It is noteworthy that working in a higher-dimensional feature space increases the generalization error of support vector machines, although given enough samples the algorithm still performs well.

III. Results:

The following tables show the results obtained and accuracy, sensitivity and specificity of the classifier for the features used.

Table 1

Classifier (kernel)	Degree	Nodes	Accuracy	Sensitivity	Specificity
SVM (linear)	**0,45,90,135**	**1,3**	**68.5**	**71.6**	**73.1**
SVM (rbf)	**0,45,90,135**	**1,3**	**75.0**	**73.7**	**78.3**
SVM (polynomial)	**0,45,90,135**	**1,3**	**79.2**	**80.2**	**82.8**

Table 2

Degree	Node	Accuracy	Sensitivity	Specificity
0	**1** **3** **5** **7**	**63.2** **66.7** **65.9** **64.7**	**65.2** **68.4** **71.5** **66.0**	**68.3** **67.2** **73.4** **65.7**
45	**1** **3** **5** **7**	**64.8** **67.3** **65.7** **68.5**	**70.8** **62.9** **65.9** **66.2**	**70.4** **63.9** **71.7** **71.9**
90	**1** **3** **5** **7**	**63.9** **65.0** **66.7** **68.1**	**66.3** **69.5** **66.3** **63.7**	**65.7** **68.8** **63.1** **64.8**
135	**1** **3** **5** **7**	**67.8** **65.9** **64.2** **64.9**	**65.9** **67.0** **69.5** **65.6**	**67.4** **65.3** **63.6** **70.1**

IV. Discussion and conclusion

This work was done to get high accuracy rate by using different texture features and different kernels of SVM classifier. The highest accuracy was found using SVM (polynomial) classifier and using the nodes at distances of all d=1,3,5,7 and was about 79.2%. This accuracy might even go much greater than this one if used with proper segmentation technique. For future work, another texture based features extraction, such as wavelet or curvelet, may be used in breast cancer classification and also by trying different types of classifiers in the purpose of improving the accuracy.

V. References

1) NehaTripathi and S. P. Panda, "A Review on Textural Features Based Computer Aided Diagnostic System for Mammogram Mass Classification Using GLCM & RBFNN," *International Journal of Engineering Trends and Technology (IJETT)* , vol. 17, no. 9, pp. 462-464, Nov. 2014

2) International Conference on Computer Science and Computational Intelligence (ICCSCI 2015) **Mammograms Classification using Gray-level Co-occurrenceMatrix and Radial Basis Function Neural Network**MellisaPratiwia, Alexandera, JeklinHarefaa, SakkaNandaa

3) **Quantitative Analysis of a General Framework of a CAD Tool for BreastCancer Detection from Mammograms**

4) https://www.mathworks.com/help/images/gray-level-co-occurrence-matrix-glcm.html

5)https://en.wikipedia.org/wiki/Support_vector_machine

6) **New Feature Extraction Method for Mammogram**

Computer Aided Diagnosis Belal K. Elfarra1 and Ibrahim S. I. Abuhaiba

7) https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html

8)https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html