INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Classification of Mammograms using Texture Features

Puneeth L.
4th Sem, M.Tech , Department of CSE, SJBIT, Bangalore, India
Dr. Krishna A. N.
Associate Professor, Department of CSE, SJBIT, Bangalore, India

Abstract:
Breast cancer is the leading cause of death among women aged 35 to 54 which gives the need of prevention of breast cancer at an early stage. Mammography is the process of detecting and screening breast cancer at an early stage and prevents death. The mammogram image has to be processed to extract the features in the image. Feature extraction is done based on the Gray-level co-occurrence matrix (GLCM). The mammogram image is the input and the classified image is the output which categorizes the image with the categories namely normal, benign and cancer. Classification is done using k-nearest neighbor classifier. The classifier calculates the distance between the query image and the images in the database and assigns class name to the query image for which the distance is least, as the output.

Keywords*: Breast cancer, Mammogram, Feature extraction, Classification*

1. Introduction

Breast cancer is one of a deadly disease which occurs mostly in women in breast tissues and it can also occur in men. Breast cancer is the formation of the malignant tumour which develops from breast cells and it cannot be avoided in any case. It should be detected at an early stage to increase the survival rate. One of the leading method for diagnosing breast cancer is screening mammography.

Mammography uses low energy x-rays usually around 30 KVp(peak kilovoltage) to examine the human breast and the cancer can be screened and it can be diagnosed using mammography. Even though in medical centers, experienced radiologists are given the responsibility of analyzing mammograms, there is always a possibility of human errors, this is the main reason why we have to analyze these images in a deeper manner, which makes these images to be processed by computers which gives more accurate results making the whole process automated. Mammography is considered as the cheapest and efficient method to detect breast cancer in the earlier stages.

Mammograms can be one of the three categories namely normal, benign and cancer. Mammograms which are considered normal are those which are not having any cancerous cells and the human is absolutely fine with no breast cancer. Mammograms which are considered benign are non-cancerous breast conditions which have many of the same symptoms as breast cancer and therefore it is difficult to tell the difference between benign and cancerous conditions from symptoms alone. Mammograms which are considered cancer are conditions which contain cancerous tissues and it requires treatment which has to be started immediately by the doctor.

Figure 1 shows a mammogram image, the black area is the image background and the white area is the dense breast tissue which tells the cancer occurrence.

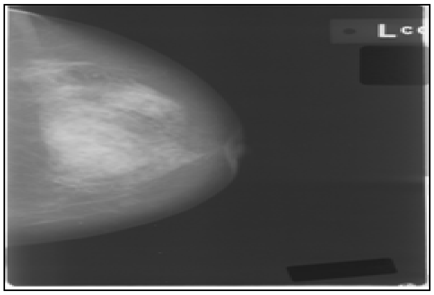


Figure 1: A Mammogram Image

Mammogram image which is used for diagnosis of breast cancer are converted into its corresponding pixel values which is representation of the image in a numerical manner. So when an image is said to be an 2-bit depth image, it contains about 4x4 values i.e., 16 values which represents the image. If an image is said to be 8-bit depth, then it contains 256x256 i.e., 65536 values which represents the image. So , we can imagine a 16-bit depth and 24-bit depth image would contain so huge values which makes it difficult to process. Therefore, lesser the bit-depth of the image, easier the process.

Basically the process of classification of mammograms includes two modules, the first module is the process of feature extraction and the second module is the process of classification. Feature extraction is done by GLCM and classification is done by using k-nearest neighbor classifier.

GLCM is the process of calculating the features which is calculated on the basis of statistical distribution of pixel intensity at a given position relative to others in a matrix of pixel which represents the image. In gray-level co-occurrence matrix, second-order statistics is used where two combinations of pixels are used to calculate the co-occurrence matrix. The combination of two pixels is considered to calculate how often the combination occurs in the image matrix which contains pixel values. This is called co-occurrence matrix. A co-occurrence matrix or co-occurrence distribution is a matrix or distribution of some values that is defined over an image to be the distribution of co-occuring values at a given offset. After the calculation of these values, these value are the input to equations which calculates entropy, uniformity etc.

The k- nearest neighbor classifier is a simple supervised classifier that has yield good performance for optimal values of k. This classifier computes the distance from the unlabeled data to every training data point and selects the best k neighbors with the shortest distance. There is no requirement for training process which actually makes this classifiers implementation as simple. The input to the classifier is the k-closest training samples and the output is the class name or an class membership. The output is decided by majority vote of its neighbors, where the input is being assigned to the class most common among its k nearest neighbors. When k is just equal to 1 , then the input is just assigned to that class.

2. Related Works

Brijeshet. al., [1] presented a system based on fuzzy neural network or neuro-fuzzy system which finds the parameters of a fuzzy system which is a learning machine by accessing the approximation techniques from neural networks and use feature extraction techniques for detecting and diagnosing micro-calcifications patterns in the digital mammograms. The neural network approach which is used for classification performs extremely well and it achieves a good classification rate with a top rate of 88.9%. This result shows how much potential the back propagation neural network i.e., BPNN has and which is used as a micro-calcification classifier. Feature vectors and neural network settings were compared together to focus the experimentation on improving the results of the classification rate which was the current interest. Since the motivation for saving human lives due to diseases like these and is inspiring researchers to develop more accurate and efficient methods of detection and diagnosis, research should continue and their strategy was to improve the classification rate up to 90 percentage.

Tippinget. al., [2] proposed an approach with a particular specialisation which is called relevance vector machine(RVM) which is identical with its functional form of support vector machine(SVM). Relevance vector machines use a Bayesian inference and obtain parsimonious solutions with machine learning techniques. RVM uses expectation maximization (EM) which are like learning method and are therefore at risk of local minima. RVM does not give global optimum which is given by SVM which uses sequential minimum optimization algorithm. The main key feature of RVM is that is can come out with a solution function that is dependent only on a very small number of training samples which are actually relevance vectors. A relevance vector machine does not need the tuning of regularization parameters during the training phase unlike SVM.

Yajieet. al., [3] presented a binary tree classifier based on the use of global features extracted from different levels of a 2-D quincunx wavelet decomposition of normal and abnormal regional images. Each leaf node of decision tree is then labelled as one of the two classes and that is where S represents the class identified as suspicious, and N presents the class identified as normal. Training set includes 120 normal and 112 abnormal are used as the test data set. The result which was obtained in this approach was fairly satisfactory since it failed to identify some of the cancers. The main reason of failure was the subtlety of some cancers and another reason was being that a tumour needs to be surrounded by the background in the block in order to be classified correctly since abnormal training and testing sets for the tree were all of this kind. The main conclusion in this approach which was given by them was to improve the efficiency of the tree classifier and reduce the clinically critical misclassification rate of abnormal regions.

Daljitet. al., [4] proposed a system which detects and does automatic classification of MRI image as well as natural images. There are basically two stages where the first stage is to do feature extraction using GLCM and PCA, and classification is done based SVM which is based on structural risk minimization where it aims on minimizing the generalized error which is errors done based on data which are unseen rather than minimizing the mean square error over the data set. This is the reason SVM tends to perform well when applied to data outside the training set. For brain MRI images, features extracted by GLCM and RBF kernel gave 100 percent accuracy whereas with PCA with the same kernel function it gave around 57.89 percent. GLCM with SVM for linear and quadratic kernel functions gives classification accuracy of 96.15%.

Saharet. al., [5] proposed a method for artificial neural network where the method is robust and effective technique, reduces computational complexity and operational time. Artificial neural network(ANNs) is a computational model inspired by animal central nervous system which is brain which is capable of machine learning as well as pattern recognition. These neural network are simply presented as systems of interconnected “neurons” which can compute values from inputs. Their work produces 92.86% for MS images with ANN classifiers.

Sahaeet. al., [5] proposed k-nearest neighbor classifier for new robust classification for technique for analyzing magnetic response images. There are three stages which they use for implementation which are feature extraction, dimensionality reduction, and classification. They use gray-level co-occurrence matrix (GLCM) to extract features from the brain MRI. K-nearest neighbor is a non-parametric method use for classification and its simplest compared to all other machine learning algorithms. The results were effective and they suggested future work on employing the proposed method on other MRI images.

3. Design and Workflow

Figure 2 is the system design or workflow which defines the conceptual structure of the whole system. The input is a mammogram image which is a gray-scale image and is converted into a pixel readable format for processing of the values. This mammogram image has to be converted from 16-bit depth or 24-bit depth to 8-bit depth. The image is used to calculate the features and then the extracted results or the values of the features of this particular image are compared with all the other images in the database. This is called the feature distance calculation. Once the distance is calculated, the image which has the least distance, that image's classification is given as the output. Every image is given with corresponding c_id in the database and this is allotted automatically based on the type of classification it belongs to. This procedure of automatically assigning the c_id which is nothing but the classification id is called supervised learning where the system learns the new data which is input into the system. The classification id and the distance of the image which has the least distance in the database is the output.

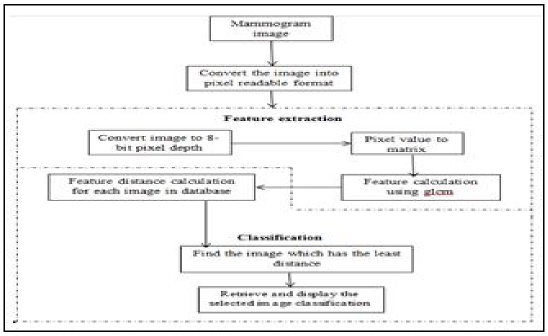
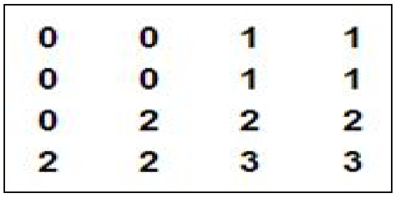

Mammogram image
Convert the image into pixel readable format
Feature extraction
Convert image to 8-bit pixel depth
Pixel value to matrix
Feature distance calculation for each image in database
Feature calculation using glcm
Classification
Find the image which has the least distance
Retrieve and display the selected im age classification

Figure 2: System Design

4. Implementation

4.1. GLCM (Gray-level Co-occurrence Matrix)

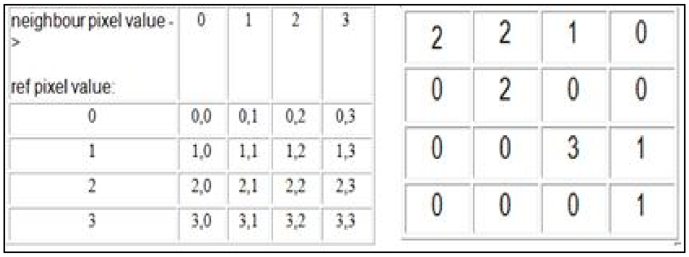
GLCM texture considers the relation between two pixels at a time, called the reference and the neighbor pixel. The neighbor pixel is chosen to be the one to the east (right) of each reference pixel.



0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

Figure 3: 4x4 matrix Test Image

Suppose considering a 4x4 matrix test image as given in figure 3 , the corresponding co-occurrence matrix has to be calculated as demonstrated in figure 4 where combination of two pixels are considered in the table and the number of times the combination occurs is calculated.



neighbour pixel value -> ref pixel value:	0	1	2	3
0	0,0	0,1	0,2	0,3
1	1,0	1,1	1,2	1,3
2	2,0	2,1	2,2	2,3
3	3,0	3,1	3,2	3,3

2	2	1	0
0	2	0	0
0	0	3	1
0	0	0	1

Figure 4: Co-occurrence Matrix

This combination is considered from left to right as well as right to left and is calculated for four angles that is 0 degree, 45 degrees, 90 degrees, 135 degrees. This is called co-occurrence matrix.Thevalues which is obtained from the co-occurrence matrix will be considered to calculate the features by using certain equations which calculates entropy, correlation, inverse difference moment, uniformity and contrast.

Entropy

$$f = -\sum_i \sum_j p(i,j) \log p(i,j)$$

Correlation

$$f = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

Inverse difference moment

$$f = \sum_i \sum_j \frac{p(i,j)}{1 + (i-j)^2}$$

Uniformity

$$f = \sum_i \sum_j \{p(i,j)\}^2$$

Contrast

$$f = \sum_i \sum_j (i-\mu)^2 \, p(i,j)$$

4.2. K-Nearest Neighbor Classifier

The k-nearest neighbor classifier computes the distance from the unlabeled data to every training data point and selects the best k neighbors with the shortest distance.Suppose, given some data instance which belongs to one of the two categories or a class, and the goal is to determine which class the new data belongs to, is the problem of classification.There is no requirement for training process which actually makes this classifiers implementation as simple.

The K-Nearest Neighbours algorithm (k-NN) is a non-parametric method used for classification. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (*k* is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour.

The input to the kNN classifier is the feature value of the query image which the user has uploaded to know the classification. Based on these input feature values, the distance is calculated for all the images from the query.Basically, classification is done by comparison of feature values or feature parameters which is got by feature extraction. When the user inputs the image for which the classification has to be done, the corresponding feature has to be extracted. After extracting the feature, feature values are calculated and those values are compared with all the other images which are already fed into the classifier. This comparison is a called distance calculation.

Feature distance is calculated by Euclidean distance between the query image which the user has input and the images which are already fed into the classifier. The distance values which are got are updated in the table and operations are done on this table to retrieve the classification.The output is the classification ID of the least distance of the whole images. This classification ID has the category name which has to be output to the user.

5. Conclusion

Classification of mammograms involves two main steps, first step is the process of feature extraction which is done based on grey-level co-occurrence matrix(GLCM) which is the second-order statistics that can be used toanalyze images as a texture. Classification is done based on three main categories namely benign, normal and cancer where classifying to the appropriate category is main problem considered here. Classification is mainly done using classifier technique called as k-nearest neighbor which assigns classification based on majority of vote of neighboring clusters

6. References

1. BrijeshVerma and John Zakos, “A Computer-Aided Diagnosis System for Digital Mammograms Based on Fuzzy-Neural and Feature Extraction Techniques,” In IEEE Transactions on Information Technology In Biomedicine, Vol. 5, No. 1, March 2001.
2. M E Tipping, “Sparse Bayesian Learning and the Relevance Vector Machine,” J. Machine Learn. Res., No. 1, Pp. 211–244, 2001.

3. Yajie sun, Charles F babbs And Edward J Delp,"Normal Mammograms Classification Based on Regional Analysis", In IEEE, Pp.375-378, 2002.
4. Daljit Singh, Kamaljeet Kaur, "Classification Of Abnormalities In Brain MRI Images Using GLCM,PCA and SVM", In International Journal Of Engineering And Advanced Technology(IJEAT) ISSN:2249-8958, Vol. 1, No. 6, August 2012.
5. SaharJafarpur, Zahra Sedghi, Mehdi ChehelAmirani, "A Robust Brain MRI For Classification With GLCM Features", International Journal Of Computer Applications, Vol. 37, No. 12, Pp. 0975-8887, January 2012.
6. A Suresh, Dr.K L Shunmuganathan, "An Efficient Texture Classification System Based on Gray Level Co-occurrence Matrix", IRACST-International Journal Of Computer Science and Information Technology & Security (IJCSITS), Vol. 2, No.4, Pp. 2249-9555, August 2012.
7. M.Pontil and A. Verri, "Support Vector Machines for 3-D Object Recognition," IEEE Transaction Pattern Analysis Machine Intelligence, Vol. 20, No. 6, pp. 637–646, Jun. 1998.
8. V Wan andW M Campbell, "Support Vector Machines for Speaker Verification and Identification," in Proc. IEEEWorkshop on Neural Networksfor Signal Processing, Sydney, Australia, pp. 775–784, Dec. 2000
9. Joshi and Phadke, "Feature Extraction and Texture Classification in MRI", A.C., Pp.975-987, 2010.
10. Zhang Y, Dong Z, Wu L and Wang SH, " A Hybrid Method For MRI Brain Image Classification", Vol. 38, Pp. 10049-10053, 2011.
11. R Randen and J H Husoy, "Filtering for Texture Classification:A Comparative Study", IEEE Transaction Pattern Analysis Machine Intelligence", Vol.21, No. 4, Pp. 291-310, April 1999.
12. R M Haralick, KShanmugam, and IDinstein, "Textural features for image classification", IEEE Transaction System, Man and Cybernetics, Vol. SMC-3, pp. 610–621, 1973.
13. R A Fisher, "The use of multiple measurements in taxonomic problems," Ann. Eugenics, Vol. 7, pp. 179–188, 1936.
14. K. Fukunaga, "Introduction to Statistical Pattern Recognition", 2nd edition, San Diego Academic, pp. 67-75, 1990.
15. S Mika, G Ratsch, J Weston, B Scholkopf and K RMuiller, "Fisher Discriminant Analysis With Kernels," in Neural Networks for Signal ProcessingIX. Piscataway, NJ: IEEE, pp. 41–48, 1999.
16. IRMA database: ganymed.imib.rwth-aachen.de/irma/datasets_en.php?SELECTED=00010