ISSN No. 0976-5697

Volume 1, No. 4, Nov-Dec 2010

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Classification of Breast Cancer Using SVM Classifier Technique

S.Srirambabu
M.S Software Engineering
School of Information Technology and Engineering
VIT University
Vellore, Tamil Nadu
Ramsrimail@yahoo.in

Santhosh Kumar. V
M.S Software Engineering
School of Information Technology and Engineering
VIT University
Vellore, Tamil Nadu
sanphite@gmail.com

B.Senthil Murugan*
Assistant Professor
School of Information Technology and Engineering
VIT University
Vellore, Tamil Nadu
senthilmurugan.b@vit.ac.in

Abstract: This paper proposes a technique for classifying the breast cancer from mammogram. The proposed system aims at developing the visualization tool for detecting the breast cancer and minimizing the scheme of detection. The detection method is organized as follows: (a) Image Enhancement (b) Segmentation (c) Feature extraction (d) Classification using SVM classifier Technique. Image enhancement step concentrates on converting an image to more and better understandable level thereby applying Median filtering approach for reducing the noise on an image. Then the Contrast stretching is applied to increase the contrast of the image. The main part of cancer detection is segmenting the breast image to improve the diagnosis and detection of breast cancer. The segmentation used here is Thresholding. Next the features are extracted from the cancer segmented area and classified the cancer according to its feature by using Support Vector Machine (SVM) classification technique. The method was tested for 119 mammogram image from the mini-mias database.

Keywords: Median filter, Contrast stretching, Segmentation, Feature Extraction, SVM, RBF, Mammogram.

I. INTRODUCTION

Breast cancer is one of the major causes for the increased mortality among women especially in developed countries [1]. Breast cancer is a malignant case that occurs in the tissue of the breast. A malignant tumor is formed from an increased number of cancer cells that develop in certain tissues and when left untreated, can spread to other areas of the body. Breast cancer can occur in both females and males, but cases of breast cancer in males are rare. Mammogram offers high quality images at low radiation doses and is the only widely accepted imaging method for routine breast cancer detection. The diagnosis result shows the cancer is of three types. (a) Normal: The mammogram display the result doesn't contain any cancerous cells of the breast. (b) Benign: The mammogram display the result contains cancerous cell consider as Non cancer. (c)Malignancy: the mammogram display the cancerous cell consider as cancer. Malignant tumor grows faster than benign and cause healthy problems [2] [3]. Computer Aided Systems which integrates computer science, image processing, pattern recognition and artificial intelligence technologies serves as a diagnosis tool for the radiologist who uses the output from the computerized analysis of medical images as a second opinion in detecting lesions and in making diagnostic decisions. These CAD systems are especially useful when the radiologist become tired of screening mammograms.

The classification scheme prescribed in this system is Support Vector Machine classifier. SVMs (Support Vector Machines) are a useful technique for data classification. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. SVM classifier has several advantages when compared to classical supervised classification methods such as maximum likelihood. SVM provides the fine class separation and allows for the use of raw image data as feature vectors, which facilitates the classification of cancer images. SVM are entitled to perform built in texture classification thereby avoiding pre-processing and supports embedded feature extraction by means of a kernel.

II. LITERATURE SURVEY

In the past several years there has been tremendous interest in image processing and analysis techniques in mammography. One common approach for detecting abnormalities in mammograms is to use a series of heuristics, e.g. filtering and thresholding which may include texture analysis to automatically detect abnormalities. Most of the schemes for early detection of breast cancer incorporate the application of techniques such as Gaussian Smoothing filtering, top hat operation and Discrete Wavelet transformation for performing the image enhancement. A Gaussian smoothing is the result of blurring an image by a Gaussian function. The function is intended to reduce the noise of an image and its detail while handling an image effect produced by an out-of-focus lens or the shadow of an object under usual illumination. Gaussian smoothing is also used as a pre processing stage in computer vision algorithms in order to enhance image structures at different scales. Gaussian smoothing is commonly used with edge detection. Smoothing an image reduces the amount of noise in an image, which allows the more prominent edges to be detected while the noisy or less prominent edges are not detected [4]. The Top-hat filter is several real-space or Fourier space filtering techniques. The name top-hat originates from the shape of the filter, which is a rectangle function, when viewed in the domain in which the filter is constructed. Fourier space is used to analyze the mathematical function with respect to frequency rather than time. Real-space form is the same as the moving average, with the exception of not introducing a shift in the output function. In numerical analysis and functional analysis, a discrete wavelet transform (DWT) is any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information (location in time). The discrete wavelet transform has a huge number of applications in science, engineering, and mathematics and computer science. Most notably, it is used for signal coding, to represent a discrete signal in a more redundant form, often as a preconditioning for data compression [5]. In this proposed work, the first step is to apply median filtering for image enhancement followed by contrast stretching. In the second stage thresholding method of segmentation is applied for segmenting the tumor portion.

III. MEDIAN FILTERING

The original mammogram is enhanced to increase the image quality and understandable level. The aim of image enhancement is to improve the interpretability or perception of information in images for human viewers, or to provide `better' input for other automated image processing techniques. Image noise is the random variation of brightness or color information in images produced by the sensor and circuitry of a scanner or digital camera. Image noise can also originate in film grain and in the unavoidable shot noise of an ideal photon detector. The technique used for image enhancement is Median filtering. It is a nonlinear neighborhood operation that can be performed for the purpose of noise reduction that can do a better job of preserving edges than simple smoothing filters. In median filtering, the neighboring pixels are ranked according to intensity value and the median value becomes the output value for the pixel under evaluation. Median filters can do an excellent job of rejecting certain types of noise, in particular, “shot” or impulse noise in which some individual pixels have extreme values.

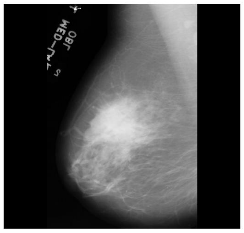


Fig.1.a Original mammogram

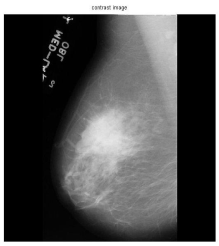


Fig.1.b Noise removed using Median Filtering

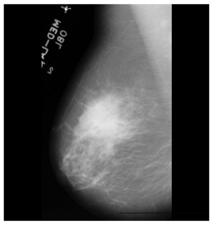


Fig.1.c Contrast Stretching

IV. THRESHOLDING

Thresholding is an important step in the cancer classification process. During the thresholding process, individual pixels in an image are marked as “object” pixels if their value is greater than some threshold value and as “background” pixels otherwise. The key parameter in the thresholding process is the choice of the threshold value. The threshhold is chosen manually and it is between 0.6 to 0.9.

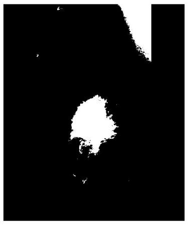


Fig.1.c Tumor Segmented Output

V. FEATURE EXTRACTION

The features used to measure the properties from the segmented image are:

Area - Scalar that specifies the actual number of pixels in the region.

Centroid - 1-by-ndims(L) vector that specifies the center of mass of the region.

ConvexArea - Scalar that specifies the number of pixels in ConvexImage. This property is supported only for 2-D input label matrices.

Eccentricity - Scalar that specifies the eccentricity of the ellipse that has the same second-moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1.

EquivDiameter - Scalar that specifies the diameter of a circle with the same area as the region.

EulerNumber - Scalar that specifies the number of objects in the region minus the number of holes in those objects.

Extent - Scalar that specifies the proportion of the pixels in the bounding box that are also in the region.

Extrema - 8-by-2 matrix that specifies the extrema points in the region. Each row of the matrix contains the x- and y-coordinates of one of the points.

FilledArea - Scalar specifying the number of on pixels in FilledImage.

MajorAxisLength - Scalar specifying the length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region.

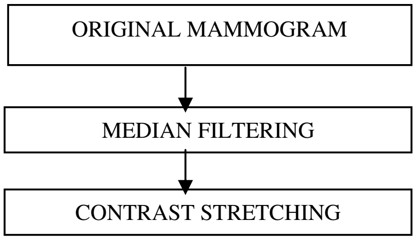
MinorAxisLength - Scalar that specifies the length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region.

Orientation - Scalar that specifies the angle (in degrees ranging from -90 to 90 degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region.

Perimeter - p-element vector containing the distance around the boundary of each contiguous region in the image, where p is the number of regions.

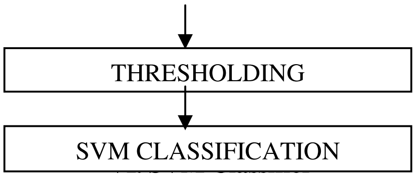
Solidity - Scalar specifying the proportion of the pixels in the convex hull that are also in the region. Computed as Area/ConvexArea.

V. BLOCK DIAGRAM


ORIGINAL MAMMOGRAM

MEDIAN FILTERING

CONTRAST STRETCHING



VI. SVM CLASSIFIER

Support Vector Machines are a set of relate supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis.The basic idea of SVM is to use linear model to implement nonlinear class boundaries through some nonlinear mapping of the input vector into the high dimensional feature space. A linear model constructed in the new space can represent a nonlinear decision boundary in the original space. In the new space, an optimal separating hyper-plane is constructed. Thus SVM is known as the algorithm that finds a special kind of linear model, the maximum margin hyper-plane. The maximum margin hyper-plane gives the maximum separation between the decision classes. The training examples that are closest to the maximum margin hyper-plane are called support vectors and the margin is the distance between the support vectors and the class boundary hyperplanes [9].All other training examples are irrelevant for defining the binary class boundaries [6]. The SVM takes a set of input data and it classifies the given tumor data sets to which type or class it belongs. More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so called functional margin). In general, larger the margin, lower the generalization error of the classifier. The hyperplanes in the large space are defined as the set of points whose cross product with a vector in that space is constant [7] [8]. With this choice of a hyperplane the points x in the feature space which are mapped into the hyperplane are defined by the relation

$$\sum_{i} \alpha_i K(x_i, x) = constant$$

Many SVM tools were considered for the sake of implementation, some of them being SVMLite, built-in SVM functions from MATLAB and LibSVM etc. The implementation phase makes use of MATLAB Support Vector Machine Toolbox. The toolbox provides routines for support vector classification and support vector regression. The prescribed environment is incorporated with the visualization tools which provided the graphical view of simple classification and regression problems. The classification approach makes use of the following parameters:

Table I. SVM Function Parameter

Parameter Name	Description
X	Training Input
Y	Training Targets
Ker	Kernel Function
C	Upper Bound(non separable case)
Nsv	Number of Support Vectors
Alpha	Lagrange Multipliers
b0	Bias Term

The maximum margin hyper-plane can be represented as the following equation in terms of the support vectors:

$$y= b+\sum \alpha_i y_i K(x(i).x)$$

The function ***K(x(i).x)*** is defined as the kernel function. There are different kernels for generating the inner products to construct machines with different types of nonlinear decision surfaces in the input space. Choosing among different kernels the model that minimizes the estimate, one chooses the best model. There are number of kernels that can be used in SVM models. These include linear, polynomial, RBF and sigmoid. The RBF is by for the most popular choice of kernel types used in SVM. There is a close relationship between SVMs and the Radial Basis Function (RBF) classifiers. The Gaussian radial basis function is given by

$$K(x,y) = \exp (- (x-y)^2 /2\boldsymbol{\sigma}^2)$$

where σ is the bandwidth of the Gaussian radial basis function kernel.

VII. RESULT

To summarize the developed method, the original mammogram image is enhanced using median filtering technique and contrast of the image increased by contrast stretching. The enhanced image is threshold to binary image using different threshold values. The features are extracted from the segmented image. Support vector classifier is used here to classify the tumor according to its features.

VII. REFERENCES

[1]. E.C.Fear, P.M.Meaney, and M.A.Stuchly,"Microwaves for Breast cancer detection", IEEE Potentials, vol.22, Pp.12-18, February-March 2003.

[2]. Homer MJ.Mammographic Interpretation: A practical Approach. McGraw hill, Boston, MA, second edition, 1997.

[3]. American college of radiology, Reston VA, Illustrated Breast Imaging Reporting and Data system (BI-RADSTM) third edition, 1998.

[4]. S.M.Astley,"Computer –based detection and prompting of mammographic abnormalities", Br.J.Radiol, vol.77, pp.S194-S200, 2004.

[5]. C. Cortes, V. N. Vapnik, "Support vector networks", Machine learning Boston, vol.3, Pg.273-297, September 1995.

[6] N. Acir, "A support vector machine classifier algorithm Based on a perturbation method and its application to ECG Beat recognition systems" Expert systems with application New York, vol.31, pg. 150-158 July 2006.

[7] O. Chapelle, V. N. Venice, Y. Bengio, "Model selection for Small sample regression", Machine Learning Boston vol.48, pg. 9-23, July 2002.

[8] Y. Liu, Y. F. Zhung, "FS_SFS: A novel feature selection Method for support vector machines", pattern recognition New York, vol.39, pg.1333-1345, December 2006.

[9]. V. N. Vapnik, "An overview of statistical learning theory", IEEE Trans. Neural Networks New York, Vol. 10, pg. 998- 999, September 19999.

[10] Y.Ireaneus Anna Rejani et al /International Journal on Computer Science and Engineering Vol.1(3), 2009, 127-130 " Early Detection of Breast Cancer using SVM Classifier Technique" .

[11] http://www.isis.ecs.soton.ac.uk/resources/svminfo/

[12] "SVM Approach to Breast Cancer Classification" by Mihir Sewak1, Priyanka Vaidya1, Chien-Chung Chan, Zhong-Hui Duan

[13] World Health Organization Fact Sheet, (2006), Cancer, http://www.who.int/mediacentre/factsheets/fs297/en/