# A COMPARISON OF DIFFERENT PATTERN RECOGNITION METHODS WITH ENTROPY BASED FEATURE REDUCTION IN EARLY BREAST CANCER CLASSIFICATION

***Liuhua Zhang***
***Wenbin Zhang***
Department of Computer Science, Memorial University of Newfoundland, Canada

**Abstract**

Breast cancer is in the most common malignant tumor in women. It accounted for 30% of new malignant tumor cases. Although the incidence of breast cancer remains high around the world, the mortality rate has been continuously reduced. Early detection by mammography is an integral part of that.In the study, we tested on three combinations of wavelet and Fourier features, including Db2, Db4, and Bior 6.8, and selected the top appropriate amounts of features which related most to the breast cancer according to the information gain. At last, three classifiers, including Back-propagation (BP) Network, Linear Discriminant Analysis (LDA), and Naïve Bayes (NB) Classifier, were tested in the original and new database, and significant figures such as sensitivity and specificity were calculated and compared.

**Keywords:** Breast cancer, feature reduction, classification

## Introduction

### Background

Breast cancer is the most commonly diagnosed form of cancer in women and the second-leading cause of cancer-related death behind lung cancer [1]. Studies show that early detection, diagnosis and therapy is particularly important to prolong lives and treat cancers. Mammography is a “specific type of imaging that uses a low-dose x-ray system to examine breasts” [2]. To date, screening mammography is the best available radiological technique for early detection of breast cancer [1].

The performance of such a mammography screening system can be measured by two parameters: sensitivity and specificity.

Sensitivity (true positive rate) is the proportion of the cases deemed abnormal when breast cancer is present. In cancer screening protocols, sensitivity is deemed more important than specificity, because failure to diagnose breast cancer may result in serious health consequences for a patient. Almost fifty percent of cases in medical malpractice relate to “false-negative mammograms” [3]. Specificity (true negative fraction) is the proportion of the cases deemed normal when breast cancer is absent. Although the consequences of a false positive, that is, diagnosing a normal patient as having breast cancer, are less severe than missing a positive diagnosis for cancer, specificity should also be as high as possible. False positive examinations can result in unnecessary follow-up examinations and procedures and may lead to significant anxiety and concern for the patient.

### Data Transformation

Fourier transform is one of the most important methods in the field of signal processing, it builds up a bridge between frequency domain and time domain.

$$F(v) = \int e^{-2\pi i(x,v)} f(x)\,dx \qquad [0.1]$$

$$f(x) = \int e^{2\pi i(x,v)} F(v)\,dv \qquad [0.1]$$

The Fourier pair illustrates that data presented in one domain may be represented in the other domain through the process of inverse transformation.

Fourier analysis is a useful tool for extracting data from many time domain signals or determining the resolution level in spatial domain images. Similarly, frequency encoded data can be transformed to the spatial domain. Perhaps the best known example of this is MRI data, collected in a frequency encoded time domain is transformed to the frequency encoded spatial domain to provide the MRI image. It can give various frequency components in the signal. However, Fourier transform has serious disadvantages: signals transformed to the frequency domain lose time information after transformation.

In 2D wavelets we have a scaling function and three wavelets.

The scaling function $\varphi^{2D} = \varphi(x)\varphi(y)$

The three wavelets $\Psi_1^{2D} = \varphi(x)\Psi(y)$

$$\Psi_2^{2D} = \Psi(x)\varphi(y)$$
$$\Psi_3^{2D} = \Psi(x)\Psi(y)$$

where ϕ and ψ indicate the scaling function and 1-D wavelet respectively. The discrete wavelet transforms of image f(x,y) of size M and N is

$$W_\varphi(j_0,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\varphi_{j_0,m,n}(x,y)$$

$$W_\varphi^i(j,m,n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\Psi_{j,\mathrm{m,n}}^i(x,y)$$

The image is broken up into a sum of orthogonal signals corresponding to different resolution scales. From the detailed coefficients we get the horizontal, vertical and diagonal detailed of the image.

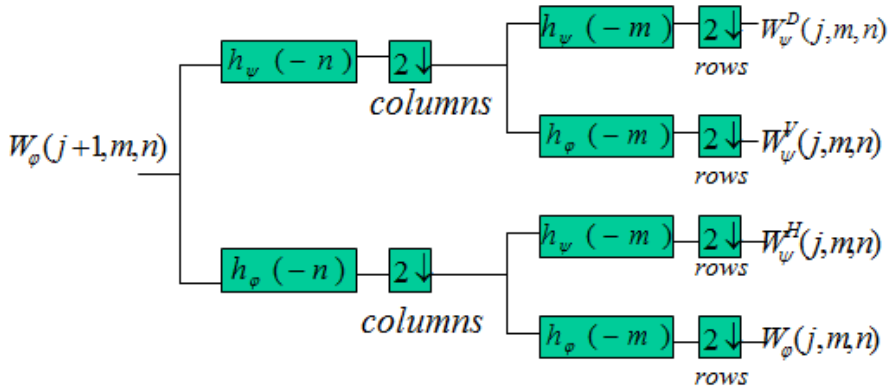


Fig. 1.1. Fast 2D wavelet transform (algorithm flow chart retrieved from the course note of Digital Image Processing in Memorial University)

Figure 1.1 illustrates the general form of 2D wavelet transform. The decompositions firstly run along the x-axis and then calculate along the y-axis, a picture then can be divided into four bands: LL (left-top), HL (right-top), LH (left-bottom) and HH (right-bottom).

**The proposed research**

**Complete image analysis system**

The entire image analysis system, from the reading of the original image to the final classification as either normal or suspicious, is represented by the block diagram of Fig. 2.1. The system consists of three distinct stages: In the image processing stage, the system inputs an original mammography image and ouputs a normalized image which could be used in

wavelet transformation. The feature selection part is mainly selecting benefit features which are extracted from Fourier and wavelet transform. The last stage is image classification, which combines different classifiers to determine whether a mammogram is normal or suspicious.

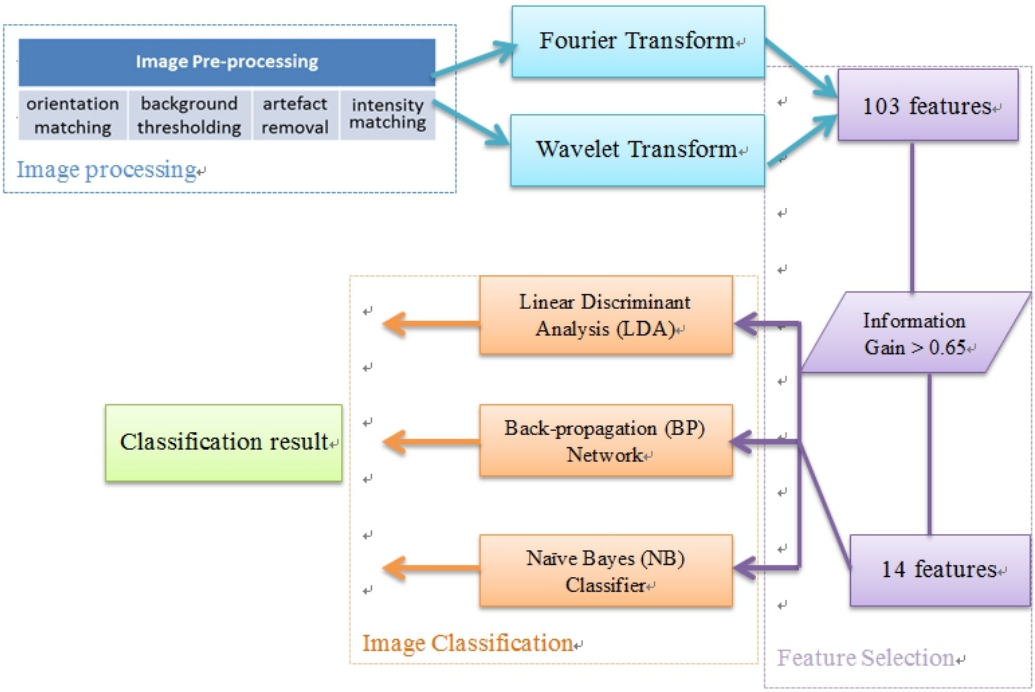


**Fig. 2.1 – Block diagram of complete image classification system**

Orientation matching step ensures that all images pointed to the same direction, preventing changes in the wavelet transform coefficients due only to the directionality change between right and left images.

Thresholding is the simplest method to create binary images, it normally sets all pixels below a set intensity level to zero [4]. A satisfied threshold can remove all irrelevant information in the background pixels, and leave foreground objects unaltered. A most commonly used method to choose the threshold is Otsu's Method, which assumes that the image to be thresholded contains two classes of pixels or bi-modal histogram (e.g. foreground and background) then calculates the optimum threshold separating those two classes so that their combined spread (intra-class variance) is minimal [5].

The thresholding process is usually implemented in conjunction with artefact removal. It is because that it was convenient to perform a binary thresholding, where all pixels below the threshold were set to an intensity of zero and all pixels above the threshold were set to an intensity of one. However, in this work, no artefact presents in mammograms, we can skip to the next step, intensity matching.

Intensity matching is the last pre-processing step that applied to the images before they are ready for wavelet decomposition. In this step, all mammograms are scaled to a relative intensity of 1.0 and all other image pixels are linearly scaled accordingly. The transformation is described by:

$$img\_out = \frac{img\_in}{\max(img\_in)}, \qquad (2.1)$$

where img_in is the input image following the background thresholding step and img_out is the intensity matched image whose pixel intensities range from zero to one. This step is proceeded to ensure the uniformity across all different images, because different mammograms could lead to different pixel intensities after coming through different machines.

**Wavelet and Fourier Transform**

After all mammograms were pre-processed from the previous steps, wavelet and Fourier transform were performed on those images. In this experiment, we will extract four statistical features: the mean intensity, the standard deviation of the pixel intensities, the skewness of the pixel intensities and the kurtosis of the pixel intensities. Further, the classification system will use some of these features to classify whole images as being normal or suspicious.

1. Mean

The mean, m of the pixel values in the defined window, estimates the value in the image in which central clustering occurs. The mean can be calculated using the formula:

$$\mu = \frac{1}{N}\sum_{i,j} I(i,j)$$

Where $I(i,j)$ is the pixel value at point $(i,j)$ of an image of size M*N.

2. Standard Deviation

The Standard Deviation, σ is the estimate of the mean square deviation of grey pixel value $I(i,j)$ from its mean value. Standard deviation describes the dispersion within a local region. It is deter-mined using the formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i,j}[I(i,j) - \mu]^2}$$

3. Skewness

The third statistic measured from each wavelet map image is the skewness of the pixel intensities. The skewness of a distribution of values is defined as the third central moment of the distribution, normalized by the cube of the standard deviation. Symbolically, the skewness S is calculated according to:

$$S = \frac{1}{N}\sum_{i,j}[\frac{I(i,j) - \mu}{\sigma}]^3$$

4. Kurtosis

The fourth and final statistic measured from the wavelet maps is the kurtosis of the pixel intensities. The kurtosis of a distribution of values is defined as the fourth central moment of the distribution, normalized by the fourth power of the standard deviation of the distribution. Symbolically, the kurtosis K is calculated according to:

$$K = \frac{1}{N}\sum_{i,j}[\frac{I(i,j) - \mu}{\sigma}]^4$$

**Feature Selection**

Feature selection is a process of reducing features according to a certain evaluation criterion, it is now normally used as a preprocessing step to machine learning. Feature selection has been researched since 1970's and it was proven to be effective in removing irrelevant and redundant features, increasing efficiency in improving learning performance like accuracy [6]. Dozens of feature selection methods have been developed during the past years and these can be divided into three categories: filter methods, wrapper methods, and hybrid methods [7]. Filter methods rely on characteristics of each individual feature using an independent test without involving any learning algorithm. Wrapper methods apply a specific machine learning algorithm and utilize its corresponding classification performance to evaluate the selected features [8]. While hybrid methods combine the advantages of filter and wrapper methods.

In this work, we evaluate the goodness of a feature according to its entropy. Entropy is a measure of the uncertainty of a random variable. The entropy of a variable X is defined as

$H(X) = -\sum_i P(x_i) log_2(P(x_i))$ (2.2)

and the entropy of X after observing values of another variable Y is defined as

$$H(X \mid Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i \mid y_j)) \qquad (2.3)$$

Where $P(x_i)$ is the prior probabilities for all values of X, and $P(x_i|y_j)$ is the posterior probabilities of X given the values of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain [9], given by

$$IG(X \mid Y) = H(X) - H(X \mid Y) \qquad (2.4)$$

If we have $IG(X \mid Y) > IG(Z \mid Y)$, it means a feature Y is regarded more correlated to feature X than to feature Z.

**Image Classification**

**Linear Discriminate Analysis**

Linear Discriminant Analysis (LDA), also called Fisher Linear Discriminant (FLD), is a classic algorithm of pattern recognition. It is introduced to the field of pattern recognition and artificial intelligence by Belhumeur in 1996 [10]. The principle idea is to project high-dimensional pattern samples in the best vector space, so that to extract the classification information and the dimension of compressed feature space. After projection, pattern samples in the new subspace have the biggest between-class distance and the minimum within-class distance, which garantee the best separability in the space. Therefore, it is an effective method for feature extraction.

Given N samples in d-dimensions, $x^{(i)}\{ x_1^{(i)}, x_2^{(i)},\ldots, x_d^{(i)} \}$. Among which, there are $N_1$ samples belonging to class $\omega_1$, other $N_2$ samples belonging to class $\omega_2$. To make sure all classes can be clearly reflected in low dimensional data, we can imagine there is a one-dimensional vector that can determine every sample's category. This best vector is named W (d-dimension), and the projection from sample X to W can be calculated as:

$$y = w^T x \qquad (2.5)$$

The value of y is the distance from the projection of sample X to the origin. When X is two dimensions, a straight line with the direction of w is needed to make projection, and the second step is to fine a straight line that can best classify sample points.

**Back-propagation Network**

BP neural network is an abbreviation for error back propagation algorithm, and it is commonly used in artificial neural network [11]. It consists of information forward propagation and error backward propagation. Shown as in Fig. 2.2, BP network is a three layer network, which includes: input layer, hidden layer and output layer.

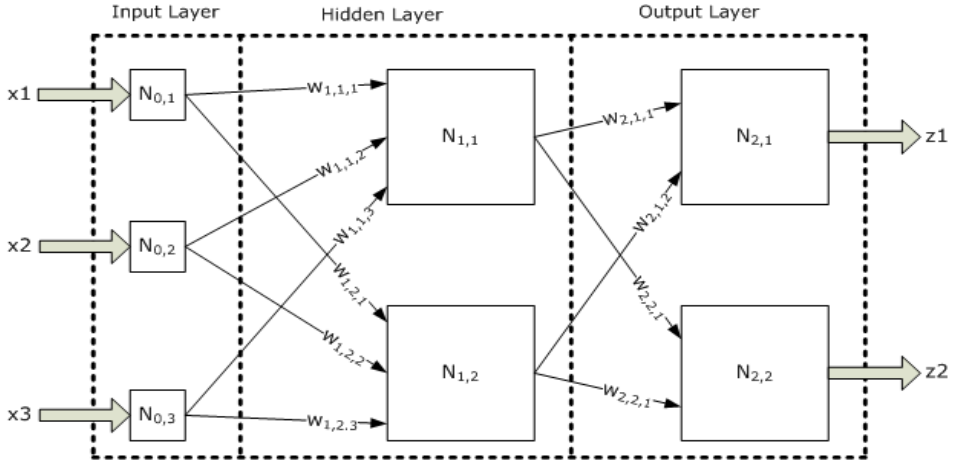


Fig. 2.2 BP neural network (retrieved from: http://www.cnblogs.com/hellope/archive/2012/07/05/2577814.html )

Neurons in the input layer are responsible for receiving outside information and transmitting to the middle layer. Hidden layer is an internal information processing layer, it can be designed as a single hidden layer or multiple hidden layer according to the demand of information transformation ability. A learning forward propagation process is completed after further processing the information of neurons from the last hidden layer to the output layer. The information processing results are obtained from the output layer.

**Naive Bayes Classifier**

Naive Bayes classifier is a supervised learning method. The classifier is first trained and concluded through the training set, and then it is used to classify the undefined data. Supposing that $A_1$, $A_2$,…$A_n$ are the features for one data set, and there are m classes, $C = \{C_1, C_2, \dots C_m\}$. Given an instance, its feature is $\{X_1, X_2, \dots X_n\}$, then the posterior probability that instance belongs to a class $C_i$ is: $P = (X \mid C_i)$. The Bayes classifier can be represented as:

$$C(X) = \underset{C_i \in C}{\arg max}\, P(C_i)\, P(X \mid C_i) \qquad (2.6)$$

It indicates that the prediction accuracy reaches the largest when instance X has the largest posteriori probability.

However, the posteriori probability is difficult to calculate in that formula, the following "Naive Bayesian hypothesis" is introduced to Naive Bayes classifier: under the given conditions of categories, all attributes $A_i$ are independent from each other. That is:

$$P(A_i|C, A_j) = P(A_i|C), \quad \forall A_i, A_j, P(C) > 0 \qquad (2.7)$$

In the Naive Bayesian classification algorithm, it can independently learn either the conditional probability that each attribute $A_i$ in the category C ($P(A_i|C)$), or the probability of each attribute $A_i$. Since the value is constant, it can be replaced with a normalization factor 'a'. Then, the posterior probability becomes:

$$P(C = c|A_1 = a_1 \dots A_n = a_n) = \alpha P(C = c) \prod_{i=1}^{n} P(A_i \mid C = c) \qquad (2.8)$$

According to formula (2.8), the optimal classification ($C = C_i$) should satisfy:

$$P(C_i| < a_1 \dots a_n >) = \frac{P(<a_1 \dots a_n > | C_i)}{P(<a_1 \dots a_n >)} P(C_i) \qquad (2.9)$$

$$P(C_i| < a_1 \dots a_n >) > P(C_j| < a_1 \dots a_n >),\ j \neq i \qquad (2.10)$$

**Results and Discussions**

**Feature selection results and discussion**

There are 102 features after preprocessing, including 6 Fourier features and 96 wavelet features. In this test, db2, db4, and bio features will be tested and compared.

After feature selecting, the program shows the information gain (IG) and entropy (E(x | y)) below.

| Feature | | E(x\|y) | IG | Order |
|---|---|---|---|---|
| Db4 | Level 3 kurtosis,h | 0.021276595744680847 | 0.6771228630838643 | 13.0 |
| | Level 3 kurtosis,v | 0.0 | 0.7008776293376933 | 14.0 |
| | Level 3 kurtosis,d | 0.0070921985815602835 | 0.6930749452746311 | 15.0 |
| | Level 4 kurtosis,h | 0.024822695035460987 | 0.6730638977527105 | 29.0 |
| | Level 4 kurtosis,v | 0.008430651599580619 | 0.6917364922566107 | 30.0 |
| | Level 4 kurtosis,d | 0.033253346635041606 | 0.6635238665784421 | 31.0 |
| | Level 5 kurtosis,h | 0.040345545216601886 | 0.655212121301062 | 45.0 |
| | Level 5 kurtosis,v | 0.028368794326241127 | 0.6689769809305656 | 46.0 |
| | Level 5 kurtosis,d | 0.024822695035460987 | 0.6730638977527105 | 47.0 |

| | | | | |
|---|---|---|---|---|
| | Level 8 mean,a | 0.03546099290780141 | 0.6607201173142287 | 84.0 |
| | Level 8 kurtosis,v | 0.021276595744680847 | 0.6771228630838643 | 94.0 |
| Fourier | std | 0.0035460992907801418 | 0.6969909224745785 | 98.0 |
| | kurtosis | 0.0 | 0.7008776293376933 | 99.0 |
| | skewness | 0.010638297872340425 | 0.6891299272077788 | 100.0 |

Table 3.1 Final results of db4 and Fourier features

| Feature | | E(x\|y) | IG | Order |
|---|---|---|---|---|
| Db2 | Level 3 kurtosis,h | 0.020163843290040204 | 0.7408423251802031 | 13.0 |
| | Level 3 kurtosis,v | 0.020895522388059702 | 0.7394641522571612 | 14.0 |
| | Level 3 kurtosis,d | 0.008955223880597015 | 0.7538821642451676 | 15.0 |
| | Level 4 kurtosis,h | 0.04179104477611941 | 0.7135477724505761 | 29.0 |
| | Level 4 kurtosis,v | 0.03210414179750289 | 0.7262091448008926 | 30.0 |
| | Level 4 kurtosis,d | 0.030977444033079803 | 0.7273358425653157 | 31.0 |
| | Level 5 kurtosis,h | 0.06679833955546789 | 0.6817697307626954 | 45.0 |
| | Level 5 kurtosis,v | 0.047029514931831266 | 0.7075223729769189 | 46.0 |
| | Level 5 kurtosis,d | 0.06381326492860222 | 0.6856599703080072 | 47.0 |
| | Level 7 mean,a | 0.036947593286811145 | 0.7206482434150133 | 68.0 |
| | Level 8 mean,a | 0.029850746268656716 | 0.7284625403297388 | 84.0 |
| | Level 8 std,a | 0.06420828359500581 | 0.6861534773199935 | 88.0 |
| Fourier | std | 0.005970149253731343 | 0.7574412275035576 | 98.0 |
| | kurtosis | 0.0 | 0.764504118933247 | 99.0 |
| | skewness | 0.011940298507462687 | 0.7503048252857403 | 100.0 |

Table 3.2 Final results of db2 and Fourier features

| Feature | | E(x\|y) | IG | Order |
|---|---|---|---|---|
| Bior 6.8 | Level 3 kurtosis,h | 0.011940298507462687 | 0.7503048252857403 | 13.0 |
| | Level 3 kurtosis,v | 0.005970149253731343 | 0.7574412275035576 | 14.0 |
| | Level 3 kurtosis,d | 0.01791044776119403 | 0.7430957207090493 | 15.0 |
| | Level 4 kurtosis,h | 0.029850746268656716 | 0.7284625403297388 | 29.0 |
| | Level 4 kurtosis,v | 0.020895522388059702 | 0.7394641522571612 | 30.0 |
| | Level 4 kurtosis,d | 0.045902817167408176 | 0.708649070741342 | 31.0 |
| | Level 5 kurtosis,h | 0.03396251865994548 | 0.723633318041879 | 45.0 |
| | Level 5 kurtosis,v | 0.03396251865994548 | 0.723633318041879 | 46.0 |
| | Level 6 kurtosis,v | 0.06122320896814013 | 0.6900103560916733 | 62.0 |
| | Level 7 mean,a | 0.04105936567809992 | 0.7150491456827223 | 68.0 |
| | Level 8 mean,a | 0.029850746268656716 | 0.7284625403297388 | 84.0 |
| | Level 8 std,a | 0.048887891794273844 | 0.7056639961144763 | 88.0 |
| Fourier | std | 0.005970149253731343 | 0.7574412275035576 | 98.0 |

| | | | | |
|---|---|---|---|---|
| | kurtosis | 0.0 | 0.764504118933247 | 99.0 |
| | skewness | 0.011940298507462687 | 0.7503048252857403 | 100.0 |

Table 3.3 Final results of bior 6.8 and Fourier features

From the three features tables, it can be seen that standard diversion, kurtosis and skewness from Fourier features were selected all the time, and their information gains were among the best of candidates. In regard to three different wavelets, the average of bior features' information gain is larger than db2 features, and db4 features' information gain is the lowest among the three. As for the four statistic features, kurtosis and skewness work the best, mean and standard diversion just appear in the higher level decomposition of db2 and bior wavelet.

**Image classification results and discussion**

Three classifiers, including Linear Discriminant Analysis (LDA), Back-propagation Network, and Naive Bayes Classifier were tested in this program. First, they were tested in the original 670 mammograms.

There could be four outcomes for a classifier in judging a sample. We named the four results in confusion matrix as true positive (TP), true negative (TN), true positive (TP), false negative (FN). "TP" means a positive instance is classified correctly as positive; "FN" refers to the positive instance wrongly classified as negative. Similarly, "TN" implies a negative instance is correctly classified as negative; otherwise it is "FP". According to these four significant figures, sensitivity ,specificity and accuracy can be calculated as the following formula:

Classification accuracy is: $acc = \frac{TP+TN}{TP+TN+FP+FT}$

Sensitivity (SN): $SN = \frac{TP}{TP+FN}$

Specificity (SP): $SP = \frac{TN}{TN+FP}$

Then, all results were shown as table 3.4.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Db4 and Fourier Features | Classifier | TP | FP | FN | TN | Sensitivity | Specificity | Accuracy |
| | LDA | 121 | 49 | 400 | 100 | 23.3% | 67.1% | 33.0% |
| | BP | 460 | 48 | 61 | 101 | 88.3% | 67.8% | 83.7% |
| | NB | 495 | 99 | 26 | 50 | 95% | 33.6% | 81.3% |

(a)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Db2 and Fourier Features | Classifier | TP | FP | FN | TN | Sensitivity | Specificity | Accuracy |
| | LDA | 176 | 23 | 345 | 126 | 33.8 % | 84.6% | 63.9 % |
| | BP | 481 | 29 | 40 | 120 | 92.3% | 80.5% | 89.7% |
| | NB | 502 | 79 | 19 | 70 | 74.9% | 47.0% | 85.4% |

(b)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Bior and Fourier Features | Classifier | TP | FP | FN | TN | Sensitivity | Specificity | Accuracy |
| | LDA | 301 | 19 | 220 | 130 | 57.8% | 87.2% | 64.3% |
| | BP | 500 | 21 | 21 | 128 | 95.9% | 85.9% | 93.7% |
| | NB | 510 | 68 | 11 | 81 | 97.8% | 54.4% | 88.2% |

(c)

Table 3.4 Sensitivity and specificity results for three classifiers

Three classifiers were further tested in 817 true positive mammograms which were achieved from the clinic. When using db4 and Fourier features, LDA classifier gave the result that 282 normals out of all true positive mammograms, BP classified 470 normal mammograms, and NB showed 495 normals. The sensitivity for these three classifiers in regard to the new true positive database is 34.5%, 57.5%, 60.6%, respectively. The sensitivity for three different features was listed in table 3.5:

| sensitivity | Db4 | Db2 | bior |
|---|---|---|---|
| LDA | 34.5% | 37.8% | 45.2% |
| BP | 57.5% | 62.4% | 70.4% |
| NB | 60.6% | 65.6% | 70.5% |

Table 3.5 Sensitivity for different classifiers regarding to different features

From the results shown as above, it can be seen that LDA classifier is more sensitivity to classify cancer, and NB gives better classification in normal mammograms. BP neural network works well in both categories. In testing the new false positive database, NB classifier has better performance, LDA works the worst. Bior features gain the highest rate, followed by db2 features. The reasons may be: 1) The false positive database contains normal mammograms actually, and NB classifier is more sensitive to classify normal ones; 2) According to the information gain after feature selection part, it can be seen bior has the highest information gain, db2 is in the second place.

**Conclusions and Future directions**

**Conclusions**

In this classification regime, we will experiment on three different classifiers using three different combinations of features. The primary objective of this research is to design a tool that combine two kinds of wavelet transform together in selecting optimal features and improve the final specificity rate. Specific research objectives are basic reached:

1. Develop a set of pre-processing steps to isolate the tissue in the images and regularize the appearance of the images to make direct comparisons possible.

This objective is successfully achieved, all mammograms are regularized and contain no artefacts after a set of propocessing steps.

2. Apply the wavelet transform and Fourier transform to parse an image and generate a set of scalar features based on the output of the transform to characterize each image.

This objective is done with some novel findings. 103 features go through a feature selection system and then around 15 features are left depending on different wavelets. After testing, their information gains are all beyond 0.65, and the highest reaches 0.76.

3. Classify the images as normal or suspicious and give the sensitivity, specificity, and accuracy of the result.

All mammograms, normal and malignant, are tested in LDA, BP, and NB classifiers. According to the results, the sensitivity and specificity can be easily calculated. It is found that LDA classifier works better to classify malignant mammogram, NB classifier achieves better performance in normal mammograms, while BP works the same in both categories.

**Future direction**

1. Fourier features can always give high information gain, but wavelet features are as appropriate. In the future work, more different wavelet features can be tested to see if they have better performance.

2. For unknown database, only true positive ones have been tested. In latter work, the being tested database can contain false negative or other mammograms.

3. The ability of classification for each classifier needs to be tested and confirmed further. If they are confirmed that some classifiers have better classification performance in specific mammogram category than all others, they can be installed in a classification system to classify that category.

**References:**

National Cancer Institute of Canada. Canadian Cancer Statistics 2004. Toronto, Canada. 2004.
American Cancer Society (2008). Global Cancer Facts & Figures. Retrieved from: http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-027766.pdf
Physician Insurers Association of America.: Breast Cancer Study. Washington, DC, Physician Insurers Association of America, 1995.
Shapiro, Linda G. & Stockman, George C. (2002). "Computer Vision". Prentice Hall. ISBN 0-13-030796-3.
Nobuyuki Otsu (1979). "A threshold selection method from gray-level histograms". IEEE Trans. Sys., Man., Cyber. 9 (1): 62–66.
Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97, 245-271.
Inza I, Larranaga P, Blanco R, Cerrolaza A-J. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 2004;31:91–103.
Langley, P. (1994). Selection of relevant features in machine learning. Proceedings of the AAAI Fall Sym-posium on Relevance. AAAI Press.
Quinlan, J. (1993).C4.5: Programs for machine learning. Morgan Kaufmann.
R.O. Duda, P.E. Hart, and D. Stork. Pattern Classification. Wiley, 2000.
Rumelhart, David E.; Hinton, Geoffrey E., Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". Nature 323 (6088): 533–536.