

Article

Feature Reduction in Graph Analysis

Rapepun Piriyakul and Punpiti Piamsa-nga *

Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Jatujak, Bangkok, 10900, Thailand; Tel.: +66-29428555 ext 1419; Fax: +66-25796245; E-mail: rapepunnight@yahoo.com

* Author to whom correspondence should be addressed; E-mail: pp@ku.ac.th

Received: 10 July 2008; in revised form: 7 August 2008 / Accepted: 7 August 2008 /

Published: 19 August 2008

Abstract: A common approach to improve medical image classification is to add more features to the classifiers; however, this increases the time required for preprocessing raw data and training the classifiers, and the increase in features is not always beneficial. The number of commonly used features in the literature for training of image feature classifiers is over 50. Existing algorithms for selecting a subset of available features for image analysis fail to adequately eliminate redundant features. This paper presents a new selection algorithm based on graph analysis of interactions among features and between features to classifier decision. A modification of path analysis is done by applying regression analysis, multiple logistic and posterior Bayesian inference in order to eliminate features that provide the same contributions. A database of 113 mammograms from the Mammographic Image Analysis Society was used in the experiments. Tested on two classifiers – ANN and logistic regression – cancer detection accuracy (true positive and false-positive rates) using a 13-feature set selected by our algorithm yielded substantially similar accuracy as using a 26-feature set selected by SFS and results using all 50-features. However, the 13-feature greatly reduced the amount of computation needed.

Keywords: Path Analysis, Graph Analysis, Feature Selection, Mammogram.

1. Introduction

Breast cancer is among the most frequent forms of cancers found in women [9]. Diagnosis of breast cancer typically includes biopsy, ultrasound, and/or imaging. Ultrasound can diagnose simple cysts in the breast with an accuracy of 96-100% [11]; however, the unequivocal differentiation between solid benign and malignant masses by ultrasound has proven to be difficult. Despite considerable efforts toward improving ultrasound, better imaging techniques are still necessary. Mammography is now commonly used in combination with computer-aided diagnosis (CAD). CAD is a computer diagnosis system to assist the radiologists in image interpretation [15]. Since the causes of some types of cancer are still unknown, it can be difficult to decide whether a tissue is cancerous or not. Currently, radiologists can refer to an automated system as a second opinion to help distinguish malignant from normal healthy tissues. An automated system can detect and diagnose probable malignancy in suspicious regions of medical images for further evaluation. Since medical images for CAD (such as X-ray, CT scan, MRI, and mammogram), include a considerable number of image features, CAD improves the detection of suspected malignancies.

Image features are conceptual descriptions of images that are needed in image processing for analyzing image content or meaning. Features are usually represented as data structures of directly extractable information, such as colors, grays, and higher derivatives from mathematical computation of the basic features such as its edges, histograms, and Fourier descriptors. Each type of feature requires a specific algorithm to process it. Therefore, only features that carry essential and non-redundant information about an image should be considered. Moreover, feature-extraction techniques should be practical and feasible to compute. Many researchers have tried to improve the accuracy of CAD by introducing more features on the assumption that this will lead to better precision. However, adding more features necessarily increases the cost and computation time.

The addition of more features does not always improve system efficiency, which has led to an investigation of feature pruning techniques [2, 3, 6, 20, 23, 30]. Foggia *et al.* [20] used a graph based method with only six features and found the performance was 82.83% true positive (TP) and 0.08% false positive (FP) per image, Fu *et al.* [13] used sequential forward search (SFS) and found that only 25 features are required, with Mean Square Error (MSE) 0.02994 by using General Regression Neural Networks (GRNN). When a support vector machine (SVM) was applied, it further reduced this to 11 features, with MSE of 0.0283.

Among the algorithms to discard non-significant features are sequential forward search (SFS), sequential backward search (SBF), and stepwise regression. SFS and SBF focus on the reduction of MSE of the detection process while stepwise regression involves both the interaction of features and the MSE value. Using stepwise logistic regression is costly since this technique is based on calculations over all possible permutations of every feature in the prediction model. These techniques use an assumption to select features that has higher relation to the classifier decision output. However, an optimal set of features must be orthogonal. With the above techniques, it is possible that information from two or more candidate features may be redundant and a feature may be dependent on another.

To improve the effectiveness of feature-discarding techniques, we propose a new method using modified path analysis for feature pruning. A weighted dependency graph of features to the output of classifier and correlation matrices among features is constructed. Statistical quantitative analysis methods (regressions and posterior Bayes) and hypothesis testing are used to determine the effectiveness of each feature in the classifier decision. Experiments are performed using 50 features found in literature and evaluate feature selection effectiveness when applied on to two learning models: ANN and logistic regression. The resulting 13-feature set is compared with prediction using all 50 original features and a 26-feature set selected by the SFS method. We found that the quality is nearly equal; however, the number of feature computations is reduced by one-half and 13/50 when compared to the 26-feature set and all-feature set, respectively.

The paper is organized as follows. Section 2 is the medical image features problems and survey on the features in medical image research. Section 3 describes the feature extraction domains. Section 4 has details of the statistical collaborative methods. Section 5 describes our proposed algorithm and section 6 is the evaluation the experiments.

2. Medical Image Feature Survey

Medical image detection from mammograms is limited to analysis of gray-scale features. Distinction between normal and malignant tissue by image density is nearly impossible because of the minuteness of the differences [20]. Thus, most feature extraction methods are extended from the derivation of limited gray scale information [1, 2, 10, 27, 30]. Medical image features can be divided into three domains: spatial, texture, and spectral. Spatial domain refers to the gray-level information in an arbitrary window size. It includes gray levels, background and foreground information, shape features, and other statistics derived from image information intensity. Texture refers to properties that represent the surface or structure of an object in reflective and transmissive images. Texture analysis is important in many applications of computer image analysis for classification, detection or segmentation of images based on local spatial variations of intensity. Spectral density or spectrum of signal is a positive real value function of a frequency associated with a stationary stochastic process, which has dimensions of power or energy. However, all useful features must be represented in a computable form.

In a previous study [12], we found that most features were extracted on the assumption that more features would enhance the detection system. There are many ways to extract new features such as modifying old features, using more knowledge from syntactic images [19], and using a knowledge base [18]. Much research has been devoted to finding the best feature or best combination of features that gives highest classification rate using appropriate classifier. Some perspectives on the situation of feature extraction and selection are reviewed next.

Fu *et al.* [13] used 61 features to select a best subset of features that produced optimal identification of microcalcification using sequential forward search (SFS) and sequential backward search (SBS) reduction followed by a General Regression Neural Network (GRNN) and Support Vector Machine (SVM). We found inconsistency between the results of the two methods *i.e.* a feature which was in the top-five most significant using the SFS but was discarded by the SBS.

Zhang *et al.* [21] attempted to develop feature selection based on the neural-genetic algorithm. Each individual in the population represents a candidate solution to the feature subset selection problem. With 14 features on their experiment, there are 2^{14} possible feature subsets. The results showed that a few feature subsets (5 features) achieved the highest classification rate of 85%. In the case of a huge number of features and mammography, however, it is very costly to select features using the neural-genetic approach.

Table 1. Feature selection and classification method from previous work.

Researcher	Domain	Features used (examples)	Classifier
Fu <i>et al.</i> [13]	Texture	Co-occurrence matrix rotation with angle 0°, 45°, 90°, 135°: Difference entropy, entropy, difference variance, contrast, angular second moment, correlation	GRNN (SFS, SBS)
	Spatial	Mean, area, standard deviation, foreground/ background ratio, area, shape moment intensity variance, energy –variance	
	Spectral	Block activity, Spectral entropy	
G. Samuel <i>et al.</i> [5]	Spatial	Volume, sphericity, mean gray level, gray level standard deviation, gray level threshold, radius of sphere, maximum eccentricity, maximum circularity, maximum compactness	Rule-based, linear discriminant analysis
E. Lori <i>et al.</i> [4]	Spatial, Patient Profile	Patient profile, nodule size, shape (measured with ordinal scale)	Regression analysis
Shiraishi <i>et al.</i> [12]	Multi Domain	Patient profile, root-mean-square of power spectrum, histograms frequency, full width at half maximum of the histogram for the outside region of the segmented nodule on the background–corrected image, degree of irregularity, full width at half maximum for inside region of segmented nodule on the original image	Linear discriminant analysis
Hening [18]	Spatial	Average gray level, standard deviation, skew, kurtosis, min-max of the gray Level, gray level histogram	SVM
Zhao <i>et al.</i> [27]	Spatial	Number of pixels, histogram, average gray, boundary gray, contrast, difference, energy, modified energy, entropy, standard deviation, modified standard deviation, skewness, modified skewness	ANN
Ping <i>et al.</i> [21]	Spatial	Number of pixels, average, average gray level, average histogram, energy, modified energy, entropy, modified entropy, standard deviation, modified standard deviation, skew, modified skew, difference, contrast, average boundary gray level	ANN and Statistical classifier

Table 1. Cont.

Songyang and Ling, [24]	Mixed features	Mean, standard deviation, edge, background, foreground-background ratio, foreground-background difference ratio of intensity, compactness, elongation, Shape Moment I-IV, Invariant Moment I-IV, Contrast, area, shape, entropy, angular second moment, inverse different moment, Correlation, Variance, Sum average	Multi-layer Neural Network
-------------------------	----------------	--	----------------------------

The Information Retrieval in Medical Applications (IRMA) [3] project used global, local, and structure features in their studies of lung cancer. The global features consist of anatomy of the object; a local feature is based on local pixel segment; and structural features operate on medical apriori knowledge on a higher level of semantics. In addition to the constraints of the global feature construction and lack of prior medical semantic knowledge, this procedure was quite difficult and costly.

The researchers' choices of medical image features depend on the objectives of the individual research. Cosit *et al.* [2], Chiou and Hwang [6], and Zoran [30] used simple statistical features on gray scale intensity, while Samuel *et al.*[5] used volume, sphericity, mean of gray level, standard deviation of gray level, gray level threshold, radius of mass sphere, maximum eccentricity, maximum circularity, and maximum compactness in their CAD system. Hening [18] used average gray scale, standard deviation, skewness, kurtosis, maximum and minimum of gray scale, and gray level histogram to identify and detect lung cancer. Shiraishi [12] studied 150 images from the Japanese Society of Radiological Technology (JSRT) database by using patient age, RMS of power spectrum, background image, degree of irregularity, full width at half maximum for inside of segment region. Lori *et al.* [4] studied on personal profile, region of interest properties, nodule size, and shape. Ping *et al.* [21] extended the new modified features, number of pixel in ROI, average gray level, energy, modified energy, entropy, modified entropy, standard deviation, modified standard deviation, skewness, modified skewness, contrast, average boundary gray level. A further investigation on using more features unrelated to medical image analysis, Windodo [23] explored fault diagnosis of induction motors to improve the feature extraction process by proposing a kernel trick. On his study, 76 features were calculated from 10 statistics in the time domain. These statistics are mean, RMS, shape factor, skewness, kurtosis, crest factor, entropy error, entropy estimation, histogram lower and histogram upper. We cannot discern their common methods of selecting features; however, we can conclude that they added more features in order to increase the efficiency of their methods. Table 1 shows a summary of the features and classifiers from previous studies.

Explorations of feature extraction analysis have been found that the effects of significant features can be direct or indirect and some features do not relate to the detection results at all. Therefore, ineffective and redundant features must be discarded.

3. Feature Domains

This section presents details on feature domains that are used for medical image classification. Generally, the original digital medical image is in the form of a gray-scale or multiple spectrum

bitmap, consisting of integer values corresponding to properties (*i.e.* brightness, color) of the corresponding pixel of the sampling grid. Image information in the bitmap is accessible through the coordinates of a pixel with row and column indices. All features that can be extracted directly using mathematical or statistical models are categorized as low-level features. High-level features are summarized from low-level features, usually by machine-learning models. Much research in medical image analysis has to deal with low-level features in order to identify high-level features. In this research, we investigate several types of low-level features in order to identify mammograms as benign or malignant. The low-level features are separated into spatial, textural, and spectral domains.

The spatial domain is composed of features extracted and summarized directly from grid information. It implicitly contains spatial relations among semantically important parts of the image. Examples of spatial features are shapes, edges, foreground information, background information, contrasts and set of intensity statistics, such as mean, median, standard deviation, coefficient of variation, variance, skewness, kurtosis, entropy, and modified moment. In this research, we also use radian of mass.

Texture features are relations among pixels in a bitmap. Representation of texture features commonly uses co-occurrence matrices to describe their properties. The co-occurrence matrix of texture describes the repeated occurrence of gray-level configuration in an image. For a texture image, $P_{\phi,d}(a, b)$, denotes the frequency that two pixels with gray levels a, b appear in the window separated by a distance d in direction ϕ .

The frequencies of co-occurrence as functions of angle and distance can be defined as:

$$\begin{aligned}
 P_{0^\circ,d}(a, b) &= | \{ [(k, l), (m, n)] \in \mathbf{D} : k-m = 0, |l-n| = d, f(k, l) = a, f(m, n) = b \} | \\
 P_{45^\circ,d}(a, b) &= | \{ [(k, l), (m, n)] \in \mathbf{D} : (k-m = d, l-n = -d) \vee (k-m = -d, l-n = d), \\
 & f(k, l) = a, f(m, n) = b \} | \\
 P_{90^\circ,d}(a, b) &= | \{ [(k, l), (m, n)] \in \mathbf{D} : |k-m| = d, l-n = 0, f(k, l) = a, f(m, n) = b \} | \\
 P_{135^\circ,d}(a, b) &= | \{ [(k, l), (m, n)] \in \mathbf{D} : (k-m = d, l-n = d) \vee (k-m = -d, l-n = -d), \\
 & f(k, l) = a, f(m, n) = b \} |
 \end{aligned}$$

where $|\{ \dots \}|$ refers to set cardinality, $f(\cdot, \cdot)$ is a gray value and $\mathbf{D} = (M \times N) \times (M \times N)$

In this paper, we take ϕ to be $0^\circ, 45^\circ, 90^\circ$, and 135° , and $d=1$. Examples of features in texture domain are:

Energy or angular second moment (an image homogeneity measure): $\sum_{a,b} P_{\phi,d}^2(a, b)$

Entropy: $\sum_{a,b} P_{\phi,d}(a, b) \log_2 P_{\phi,d}(a, b)$

Maximum probability: $\max_{a,b} \{ P_{\phi,d}(a, b) \}$

Contrast: $\sum_{a,b} |a - b|^k P_{\phi,d}(a, b)$

Inverse difference moment: $\sum_{a,b,a \neq b} \frac{P_{\phi,d}^{\lambda}(a, b)}{|a - b|^k}$

Correlation (a measure of image linearity, linear direction structures in direction ϕ)

$$\sum_{a,b,a \neq b} \frac{[(ab)P_{\phi,d}(a, b)] - \mu_x \mu_y}{\sigma_x \sigma_y} \text{ where } \mu_x, \mu_y, \sigma_x, \sigma_y \text{ are means and standard deviations.}$$

Spectral features [3] are used to describe the frequency characteristics of the input image. The features are based on transformation from the spatial and time domains. Most frequently-used spectral features are based on discrete cosine transform (DCT) and wavelets. Examples of features based on the frequency domain are:

$$\text{Spectral entropy: } -\sum_i \sum_j \bar{X}(i, j) h(\bar{X}(i, j))$$

$$\text{Block activity: } A = -\sum_i \sum_j |\bar{X}(i, j)| \text{ where } i, j \text{ are window size and } \bar{X}(i, j) = \frac{|\bar{X}(i, j)|}{A}$$

The above features are frequently found in the literature of medical image analysis; there are many more features available.

4. Methodology

We hypothesize that using only one statistical method for classification will not be successful because of the restriction on measurement values of features and output. As this restriction, we investigate statistical techniques to fulfill the feature selection process. These statistical techniques consist of four parts: 1) feature classification, 2) path analysis, 3) exploration on relations among features and outputs, and 4) hypothesis testing. In the feature classification, we use correlation analysis to transform a number of features into a number of groups. In path analysis, the conceptual relations among different feature classes are constructed. Then, relations among features and between features and outputs are determined by three methods: logistic regression, simple regression, and multiple regression. Finally, hypotheses of feature relationships are tested by a Bayesian technique.

4.1. Feature classification

Since most low-level features are extracted from spatial and texture based, which are highly correlated, the feature selection strategy is subject to this limitation. The correlation coefficient is used to analyze these features. The correlation coefficient p between random variables x and y is defined as

$$p(x, y) = \frac{\text{cov}(x, y)}{\sqrt{V(x)V(y)}} \text{ where } \text{cov}(x, y) \text{ denotes the covariance of } x \text{ and } y, V(x) \text{ and } V(y) \text{ are variances}$$

of x and y . p is between -1 and 1, and $p = 0$ indicates no linear relation between x and y .

Correlation coefficients of features can be used to classify many highly related features into groups.

4.2. Path analysis

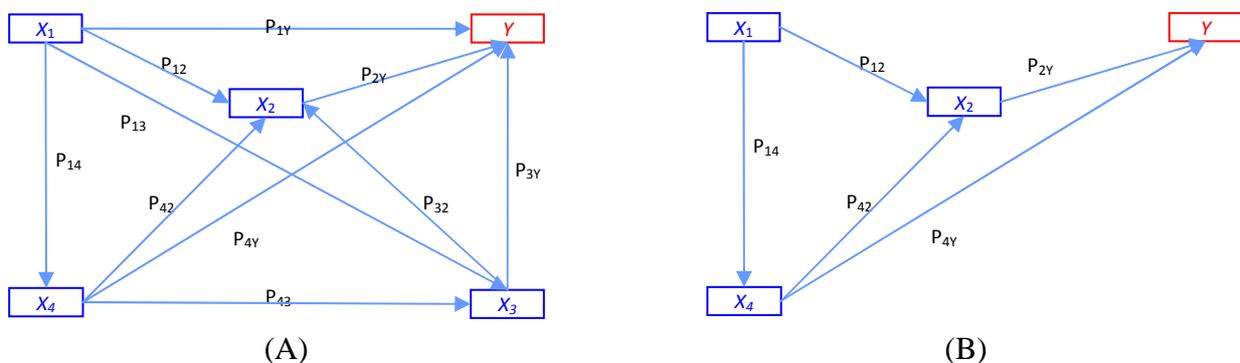
By the previous phase, we can identify groups of highly-related features. We find that the relationships of features within each group and relationships among groups to final output can be determined by path analysis.

Path analysis utilizes multiple regression analysis. Regression analysis is an analysis of causal models when single indicators are endogenous variables of the model. In a path model, there are two types of variables: exogenous and endogenous. Exogenous variables may be correlated and may have

direct effects as well as indirect effects on endogenous variables. Causality is a relationship between an exogenous variable and endogenous variable(s); philosophical causation refers to the set of all particular “causal” relations.

Being a regression-based technique, path analysis is limited by the requirement that all variables be continuous. Because our study involves continuous cause variables while the endogenous output variable is dichotomous (discrete), we cannot use path analysis directly; however, the analysis is still a graph-based process. Causal relation analysis can be explained by dependent variables that are measured on an interval or ratio scale [17]. Thus, for path analysis involving continuous endogenous variables, the categorical endogenous might have difficulty both in theoretical terms and prediction implication. Goodman [9] considered path analysis of binary variables by using logistic regression. Hagenars [10] made a general discussion of path analysis of recursive causal systems of categorical variables by using the directed log-linear model approach, which is a combination of Goodman’s approach and graphical modeling. Example of the different models of trait effects on output y is illustrated in Figure 1. Figure 1A shows a multiple regression model where each trait operates simultaneously on fitness y . Figure 1B is the path analysis model showing four traits at four time periods.

Figure 1. An example of a general recursive causal system with four independent features and a dependent output. (A) Illustration of possible relations among features and output. (B) The result of feature selection by analogy with graph base.



A path diagram not only shows the nature and direction of causal relationships but also estimates the strength of relationships. Comparatively weak relationships can be discarded; thus some features are eliminated. A path coefficient is the standardized slope of the regression model. This standardized coefficient is a Pearson product – moment correlation. Basically, these relationships are assumed to be unidirectional and linear. To overcome this limitation, we use regressions and Bayesian inference to construct a graphical model.

4.3. Relations among features and outputs

From the previous details about features and the path analysis, it is necessary to explore the cause and effect features by regression analysis. In our purpose, we suggest to use logistic regression, simple regression, and multiple regressions.

- a) Using logistic regression. Logistic regression is a regression model for Bernoulli-distributed dependent variables. It is a linear model that utilizes the logit as its link function. Logistic regression has been used extensively in medical and social sciences [4, 11]. The logit model takes the form: $\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + e_i; i = 1, 2, \dots, n$,

where $p_i = \Pr(y_i=1)$, $\beta_j > 0$; $j = 1, 2 \dots k$ are parameters (weight) of feature x_i and e_i is a random error (bias) of feature vector of a sample data.

Logistic regression model can be used to predict the response features to be 0 or 1 (benign or malignant in the case of mammogram detection). Rather than classifying an observation into one group or the other, logistic regression predicts the probability p of being in either group. The model predicts the log odds ($p/(1-p)$) that an observation later be transformed to p as value of 0 or 1 with an optimal threshold. The general prediction model is $\log(p/(1-p)) = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where \mathbf{x} is feature vector; $\boldsymbol{\beta}$ is a parameter vector; and $\boldsymbol{\epsilon}$ is a random error vector.

- b) Using simple regression and multiple regression. Simple regression has the same basic concepts and assumptions as logistic regression but the dependent variable is continuous and the model has only a single independent variable. The simple regression can be modeled as $Y_i = \beta_0 + \beta_1 X_{1i} + e_i$; $i = 1, 2, \dots, n$ where Y_i is the dependent variable, β_0, β_1 are parameters (weights), and n is the size of training data. X_{1i} is an explained variable of data record i and e_i is a random error. Regression yields a p value for the estimator of β_1 that can be used to decide whether Y has a linear relation to X . Multiple regression is an extension of simple regression model to multiple variables.

Simple logistic regression and multiple logistic regression are used to explore the cause features to effect output.

4.4. Hypothesis testing

Although the statistical techniques in previous Section can be used to identify causal features, they cannot classify those features as direct or indirect. We use hypothesis testing for this.

An appropriate way to test the hypothesis about the direction of causal relationships is easier to illustrate an abstract concept by analogy with Bayesian inference. Bayesian inference uses the scientific method, which involves collecting evidence that may or may not be relevant to a given phenomenon. The more evidence is accumulated, the degree of belief in a hypothesis changes. With enough evidence, the degree of belief will often become very high or very low. It can be used to discriminate conflicting hypotheses. Bayesian inference usually relies on degrees of belief, or subjective probabilities. Bayes's theorem adjusts probabilities based on new evidence as $P(H_o | E) = \frac{P(E | H_o)P(H_o)}{P(E)}$, where H_o represents the hypothesis; $P(H_o)$ is the prior probability of

H_o ; $P(E/H_o)$ is the conditional probability of availability the evidence E given that the hypothesis H_o is true; and $P(E)$ is the marginal probability of E , which is the probability of witnessing the new evidence E under all mutually exclusive hypotheses. $P(E/H_o)$ is the posterior probability of H_o given E .

Using hypothesis testing on the regression, we can use path analysis for the discrete output.

5. Proposed Algorithm

To solve this solution, simple regression, logistic regression, and Bayesian inference take into account of causality extraction problem. The algorithm is described as following steps.

Step 1: Partition the original feature sets ($\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_n$) into subsets using coefficients of the correlation matrix. Let the feature subsets be $S_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i} \dots \mathbf{x}_{ji})$, $i=1, 2 \dots k$ with p_{ij} being the correlation coefficient between \mathbf{x}_i and \mathbf{x}_j .

This step is to partition all features into feature subsets S_i , where S_i and S_j ($i \neq j$) are lowly dependent based on the correlations.

Step 2: Perform simple logistic regression of each independent feature $\mathbf{x}_{ji} \in S_i$, $j=1, 2 \dots R_i$ and dependent output y and then select the possible solution which satisfies a threshold value P .

The result from this step is a subset $A_i = (\mathbf{x}_{ri}, \mathbf{x}_{pi} \dots \mathbf{x}_{ki})$ of features from S_i is where each element of A_i is a direct causal feature of output y .

Step 3: Perform multiple logistic regression by using all features in set S_i , $i=1, 2 \dots k$ in the model and selecting the signified features $B_i = (\mathbf{x}_{ti}, \mathbf{x}_{li} \dots \mathbf{x}_{zi})$ from the model, where B_i is a set of direct features and indirect cause features.

Step 4: Let $D_i = A_i \ominus B_i$; where \ominus is our testing hypothesis operator for exploring the causal relations using the Bayesian inference conceptual framework.

This step is performed using Bayesian inference as in the following example for two features:

$$\text{If feature } \mathbf{x}_{ni} \text{ is the cause of } y \approx P(y/ \mathbf{x}_{ni}) > C \tag{1}$$

$$\text{If feature } \mathbf{x}_{ti} \text{ is related (highly correlated) to } \mathbf{x}_{ni} \approx P(\mathbf{x}_{ni}, \mathbf{x}_{ti}) > C \tag{2}$$

$$\text{If feature } \mathbf{x}_{ti} \text{ is not significant to } y \approx P(y/ \mathbf{x}_{ti}) < C \tag{3}$$

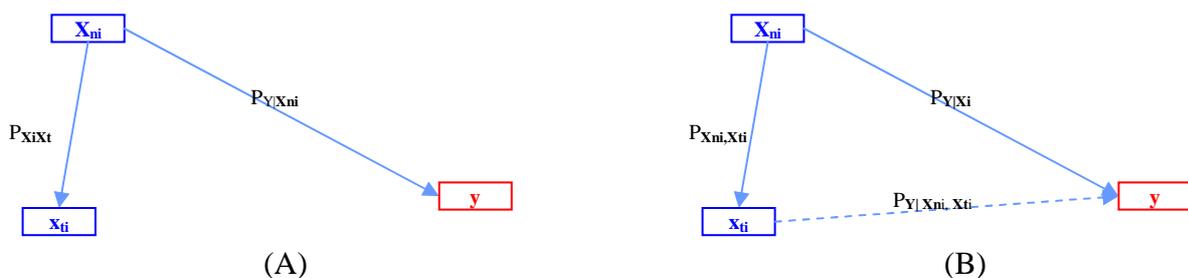
$$\text{If features } \mathbf{x}_{ni} \text{ and, } \mathbf{x}_{ti} \text{ are significant to } y \approx P(y/ \mathbf{x}_{ni}, \mathbf{x}_{ti}) > C \tag{4}$$

where C is a given threshold.

This step iteratively refines the search for the indirect cause feature with the highest correlation with the direct cause \mathbf{x}_{ni} .

Through the above predicates (1) to (4), we can accept the *hypothesis* that \mathbf{x}_{ni} and the combination of \mathbf{x}_{ni} and \mathbf{x}_{ti} cause y . Figure 2 illustrates the relations among \mathbf{x}_{ni} , \mathbf{x}_{ti} , and y .

Figure 2. The connected graph on two cause features and effect y . There is no direct effect of feature \mathbf{x}_{ti} on y in (A) but, as shows in (B), there is an interaction effect of feature \mathbf{x}_{ti} in addition with \mathbf{x}_{ni} on y .



Step 5: Repeat from Step 2 while $i \leq k$. This step produces sets D_i , where $i=1, 2 \dots k$. Note that some of D_i may be null sets.

Step 6: Construct graph G by merging subgraphs D_i ; $i=1, 2 \dots k$;

$G(V, E | Y) = \cup_{i=1}^k D_i$; $V = (v_i)$; $E = (e_i)$; Y is the effect or dependent vertex.

6. Experiment and Results

6.1. Experiment

Our experiment is based on a training set of 113 ROIs from the Mammographic Image Analysis Society (MIAS) mammogram images that are segmented by radiologists. After image segmentation, 50 features from the spatial, texture, and spectral domains are extracted. The feature set consists of mass radian, mean, maximum, median, standard deviation, skewness, kurtosis of gray level from spatial domain, energy, entropy, modified-entropy, contrast, inverse different moment, correlation, maximum, SD_x (standard deviation) and SD_y from the co-occurrence matrix of gray scale used $P_{\phi,d}(a,b)$ with distance $d=1$ and angle $\phi = 0^\circ, 45^\circ, 90^\circ, 135^\circ$ from texture domain and block activity, spectral entropy from the spectral domain. Step 1 of the experiment is to classify homogeneous features into 12 feature sets, using the bivariate correlation coefficient. Table 2 shows list of features in each set.

Table 2. Partition of the 50 original features into 12 feature sets.

Feature set	Number of features	List of Features
#1	4	Entropy rotations from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#2	4	Energy rotations from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#3	4	Inverse difference Moment rotations from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#4	4	Mean Co-occurrence rotations from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#5	4	Max Co-occurrence rotations from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#6	4	Contrast rotations from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#7	4	Homogeneity rotations from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#8	4	Standard deviations on X rotation from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#9	4	Standard deviations on Y rotation from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#10	4	Modified entropy rotations from $0^\circ, 45^\circ, 90^\circ, 135^\circ$
#11	7	mean, maximum, median, standard deviation (SD), coefficient of variation (CV), skewness, kurtosis (intensity of gray level)
#12	3	block activity, spectral entropy, mass radian

After Step 1, the simple and multiple logistic regression analysis in each feature set are performed. Tables 3 and 4 illustrate example results from Step 2 to Step 4 by using features in feature set #1.

Table 3. The effects among features in feature set #1.

Relations in Feature set #1	Effects of dependent features (using simple linear regression)
Entropy 0° to Entropy 45°	0.000 **
Entropy 0° to Entropy 90°	0.004 *
Entropy 0° to Entropy 135°	0.000 *
Entropy 45° to Entropy 90°	0.000 **
Entropy 45° to Entropy 135°	0.022 *
Entropy 90° to Entropy 135°	0.000 **

* denotes significant with 5% threshold and ** denotes highly significant with 1% threshold.

Table 3 shows the effects among features in set #1. Values in Table 3 are used to test null hypotheses that two testing features are not correlated. If any effects that have p -value less than 0.05, those pairs of features are accepted as correlated.

Tables 3 and 4 show that:

- From Table 3: Entropy 0° and Entropy 45° are highly significantly related.
- From the second column of Table 4: based on the simple logistic model, only Entropy 0° causes y (Entropy 0° is significant to y).
- From the third column of Table 4: on the multiple logistic regression model, Entropy 0° and Entropy 45° cause y .
- Finally, with Bayes inference, the direct effect is Entropy 0° and the indirect effect is the interaction of Entropy 0° and Entropy 45° cause y .

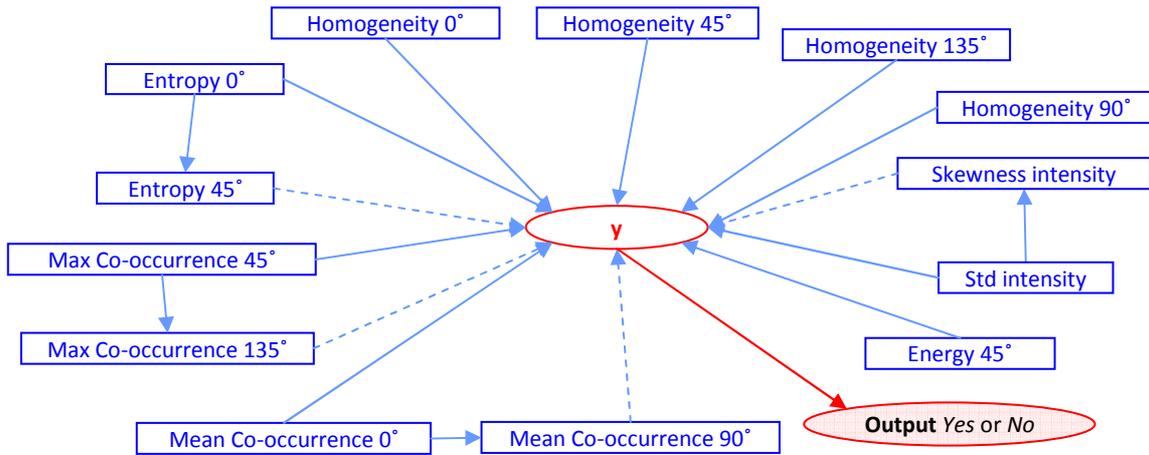
Table 4 shows the result of Step 4, $D_i = A_i \ominus B_i$ where $i=1$. After k iterations of the algorithm, the experiment results in the number of features being reduced from the original 50 to 13 features. Those features are Entropy 0°, Entropy 45°, Max Co-occurrence 45°, Max Co-occurrence 135°, Mean Co-occurrence 0°, Mean Co-occurrence 90°, Energy 45°, Homogeneity 0°, Homogeneity 45°, Homogeneity 90°, Homogeneity 135°, Standard deviation and Skewness of intensity value. The constructive cause and effect graph, $G(V,E/y)$, is shown as Figure 3.

Table 4. The effects of features in feature set #1 on output.

Feature set #1	Effects on output	
	Using simple logistic regression	Using multiple logistic regression
Entropy 0°	0.034 *	0.026 *
Entropy 45°	0.433	0.031 *
Entropy 90°	0.363	0.241
Entropy 135°	0.159	0.169

* denotes significant with 5% threshold and ** denotes highly significant with 1% threshold.

Figure 3. Complete graph on the experiment with direct and indirect effect from retaining process. (Dotted lines show indirect effects).



6.2. Verification

The effectiveness of our selected 13-feature set (our-13) is compared to the results of the all-feature set (all-50) and 26-feature set from SFS (SFS-26) on two learning systems: ANN and logistic regression. True positive (TP), false positive rate (FP) and minimum squared error (MSE) are metrics in the comparison. Tables 5 and 6 show the results from ANN and logistic regression, respectively. Both tables show that the effectiveness of our-13 is better than of SFS-26 and it is much closer to all-50. This shows that our method can detect comparably the same results while the feature computation is reduced by half compared to SFS and 13/50 compared to using all features.

Table 5. Performance of logistic regression using all-50, SFS-26 and our-13 feature sets.

Logistic regression	TP (%)	FP (%)	MSE
Using original 50 features (all-50)	82.94	14.51	0.052
Using selected 26 features (SFS-26)	77.41	18.72	0.102
Using selected 13 features (our-13)	81.64	15.06	0.084

Table 6. Performance of ANN using all-50, SFS-26, and our-13 feature sets.

ANN	TP (%)	FP (%)	MSE
Using original 50 features (all-50)	83.32	14.42	0.034
Using selected 26 features (SFS-26)	78.59	16.02	0.083
Using selected 13 features (our-13)	82.35	15.02	0.065

6.3. Analysis of results

Graph-based analysis was examined using statistical techniques to identify the crucial direct or indirect features for breast cancer detection in medical images. Our algorithm requires time complexity $O(n^2)$. We can accept the hypothesis that there is no significance between 50 features and 13 features for ANN and logistic regression with threshold 5%. A comparison of the performance between the

different configurations of architectures over two set of features (50 and 13 features) with two classifiers (ANN and logistic regression) indicates that the selected 13 features provide the best results in terms of precision with respect to computation time. Using our approach, the detection step improves the temporal ratio of computation by number of features by 50:13. Moreover, the proposed method demonstrates satisfactory performance and cost compared to SFS.

In our experiment, the 50 features were partitioned into 12 feature sets with S_{11} being the largest set. With this set, the search space for direct cause features (A_7) is $({}^7C_1)$ while indirect cause (B_7) exploration was $({}^7C_i)$ $i=2, 3 \dots 7$. We also found that there were 11 features from the texture domain and two features from the spatial domain that were eliminated from the selection process. The mass radian was not a significant feature because some masses on benign images were larger than on malignant images. Instead of using mass radian (microcalcification), the distribution of microcalcification is more advantageous.

On the theoretical aspect of finding a best combination feature set, the only way to guarantee the selection of an optimal feature set is an exhaustive search of all possible subsets of features. However, the search space could be very large: 2^N for a set of N features. Our algorithm provides a divide and conquer strategy; with N features (assume that there are r groups with k features each), the number of possible subsets for examining the feature selection is $r^k C_i$; $i=1, 2 \dots k$.

7. Conclusions

In this research, a method to reduce a number of features for medical image detection is proposed. We use mammograms from the Mammographic Image Analysis Society (MIAS) as test data and applied the proposed algorithm to reduce the number of features from a frequently-used 50 features to 13 features, while the accuracies using two learning models are substantially the same. Our method can reduce the computation cost of mammogram image analysis and can be applied to other image analysis applications. The algorithm uses simple statistical techniques (path analysis, simple logistic regression, multiple logistic regressions, and hypothesis testing) in collaboration to develop a novel feature selection technique for medical image analysis. The value of this technique is that it not only tackles the measurement problem by path analysis but also provides a visualization of the relation among features. In addition to ease of use, this approach effectively addresses the feature redundancy problem. The method proposed has been proven that it is easier and it requires less computing time than using SFS, SBF and genetic algorithms. For further research, a deeper analysis of the texture domain and the dispersion of microcalcification may provide a more efficient breast CAD system, with cost reduction and higher precision.

Acknowledgements

This research is partially supported by the Kasetsart University Research and Development Institute. Authors would like to thank Nutakarn Somsanit, MD, of Rajburi Hospital for her advice about the training data. Lastly, authors also would like to thank Dr. James Brucker of the Department of Computer Engineering, Kasetsart University for his comments on writing.

References:

1. Hiroyuki, A.; Herber, M.; Junji, S.; Qing, L.; Roger, E.; Kunio, D. Computer – Aided Diagnosis in Chest Radiography: Results of Large –Scale Observer Tests at the 1886-2001 RSNA Scientific Assemblies. *Radiographics* **2003**, *23*, 255-265.
2. Cosit, D.; Loncaric, S.L. Ruled Based Labeling of CT head Image. *Proc. 6th Conference on Artificial Intelligence in Medicine, Europe 1997*; pp. 453-456.
3. Gliman, D.M.; Sizzanne, L. State of the Art FDG Pet Imaging of Lung Cancer. *Semin. Roentgenol.* **2005**, *40*, 143-153.
4. Dodd, L.E.; Wagner, R.F.; Armato, S.G.; McNitt-Gray, M.F.; Beiden, S.; Chan, H.P.; Gur, D.; McLennan, G.; Metz, C.E.; Petrick, N.; Sahiner, B.; Sayre, J. Assessment Methodologies and Statistical Issues for Computer-Aided Diagnosis of Lung Nodules in Computed Tomography. *Acad. Radiol.* **2004**, *11*, 462-474.
5. Almato, G.S.; Roy, A.S.; MacMahon, H.; Li, F.; Doi, K.; Sone, S.; Altman, M.B. Evaluation of Automated Lung Nodule Detection on Low dose Computed Tomography Scan From a Lung Cancer Screening Program. *AUR. Acad. Radiol.* **2005**, *12*, 337-346.
6. Chiou, G.I.; Hwang, J.-N. A Neural Network Based Stochastic Active Nodule (NNS-SNAKE) for Contour Finding of Distinct Features. *Image Process. IEEE Trans.* **1995**, *4*, 1407-1416.
7. Goodman L.A. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **1971**, *61*, 215-231.
8. Hagenarrs, J.A. Categorical causal modeling latent class analysis and discrete log-linear models with latent variables. *Sociol. Methods Res.* **1998**, *26*, 436-486.
9. *Lung Cancer Home Page* (http://www.lungcancer.org/patients/fs_pc_lc_101.htm; accessed December 25, 2007)
10. Guler, I.; Ubeyli, E.D. Expert systems for time-varying biomedical signals using eigen vector methods. *Expert Syst. Appl.* **2007**, *32*, 1045-1058.
11. Song, J.H.; Venkatesh, S.S.; Conant, E.A.; Arger, P.H.; Sehgal, C.M. Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-Aided Diagnosis of Breast Masses. *Acad. Radiol.* **2005**, *12*, 487-495.
12. Shiraishi, J.; Abe, H.; Li, F.; Engelmann, R.; MacMahon, H.; Doi, K. Computer-aided Diagnosis for the Detection and Classification of Lung Cancers on Chest Radiographs. Science Direct. *Acad. Radiol.* **2006**, *13*, 995-1003.
13. Fu, J.C.; Lee, S.K.; Wong, S.T.C.; Yeh, J.Y.; Wang, A.H.; Wu, H.K. Image segmentation feature selection and pattern classification for mammographic microcalcifications. *Comput. Med. Image* **2005**, *29*, 419-429.
14. Joreskog, K.G.; Sorbom, D. *LISREL 7 User's Reference Guide*; SPSS Inc.: Chicago, 1989.
15. Doi, K.; MacMahon, H.; Katsuragawa, S.; Nishikawa, R.M.; Jiang, Y. Computer-aided diagnosis in radiology: Potential and pitfall. *Eur. J. Radiol.* **1999**, *31*, 97–109.
16. Zhao, L.; Boroczky, L.; Lee, K.P. False positive reduction for lung nodule CAD using support vector machines and genetic algorithms. *Comput. Assist. Radiol. Surg.* **2005**, *1281*, 1109-1114.

17. Lehmann, T.M.; Guld, M.O.; Thres, C.; Fischer, B.; Spitzer, K. Content-Based Access to Medical Images; available online: http://phobos.imib.rwth-aachen.de/irma/ps-pdf/MI2006_Resubmission2.pdf.
18. Miller, H.; Marquis, S.; Cohen, G.; Poletti, P.A.; Lovis, C.; Geissbuhler, A. *Automatic: Abnormal Region detection in Lung CT images for Visual Retrieval*. University and Hospital of Geneva, Service of Medical Informatics, Department de Radiologie et Informatique Medicale Home Page. <http://www.simhcuge.ch/medgift> (accessed September 5, 2007)
19. Pietikainen, M.; Ojala, T.; Xu, Z. *Rotation Invariant Texture Classification Using Feature Distributions*; available online: www.mediateam.oulu.fi/publications/pdf/7.
20. Foggia, P.; Guerriero, M.; Percannella, G.; Sansone, C.; Tufano, F.; Vento, M. A Graph-Based Method for Detecting and Classifying Clusters in Mammographic Images. *Lect. Notes Comput. Sci., Struct. Syntact. Stat. Patt. Recog.* **2006**, *4109*, 484-493.
21. Zhang, P.; Verma, B.; Kumar, K. Neural Vs Statistical Classifier in Conjunction with Genetic Algorithm Feature Selection in Digital Mammography. In *Proc. 2004 IEEE Int. Joint Conf. Neural Networks* **2004**, *3* (25-29), 2303 – 2308.
22. Jiang, W.; Li, M.; Zhang, H.; Gu, J. Online Feature Selection Based on Generalized Feature Contrast Model. In *IEEE Int. Conf. Multimedia Expo (CME)*, 2004.
23. Widodo, A.; Yang, B.-S. Application of nonlinear feature extraction and support vector Machines for fault diagnosis of induction motors. *Expert Syst. Appl.* **2007**, *33*, 241-250.
24. Songyang, Y.; Ling, G. A CAD System for the Automatic Detection of Clustered Microcalcifications in Digitized Mammogram Films. *IEEE T. Med. Imaging* **2000**, *19*, 115-126.
25. Yang, B.S.; Han, T.; Hwang, W. Application of multi-class support vector rotating machinery. *J. Mech. Sci. Tech.* **2005**, *19*, 845-858.
26. Chiou, Y.; Lure, Y. Ligomenides. Neural network image analysis and Classification in hybrid lung nodule detection (HLND) system. In *Proc. IEEE-SP Workshop Neural Networks Signal Process. 1993*; pp. 517-526.
27. Zhao, W.; Yu, X.; Li, F. Microcalcification Patterns Recognition Based Combination of Auto association and Classifier. *Lect. Notes Comput. Sci., Comput. Intell. Secur.* **2005**, *3801*, 1045-1050.
28. Zheng, B.; Qian, W.; Clarke, L.P. Digital mammography: mixed feature neural network with spectral entropy decision for detection of microcalcifications. *IEEE T. Med. Imaging* **1996**, *15*, 589-97.
29. Liang, Z.; Jaszczak, R.J.; Coleman, R.E. Parameter Estimation of Finite Mixtures Using the EM Algorithm and Information Criteria with Application to Medical Image Processing. *IEEE T. Nucl. Sci.* **1992**, *39*, 1126 - 1133.
30. Majcenic, Z.; Loncaric, S. *CT Image Labeling Using Simulated Annealing Algorithm*. (<http://citeseerx.ist.psu.edu>; accessed July 12, 2007).