A LOGISTIC REGRESSION APPROACH TO CONTENT-BASED MAMMOGRAM RETRIEVAL

Chia-Hung Wei and Chang-Tsun Li

Department of Computer Science
University of Warwick
Coventry CV4 7AL,UK
{rogerwei, ctli}dcs.warwick.ac.uk

ABSTRACT

Content-based image retrieval (CBIR) has been proposed to address the problem of image retrieval from medical image databases. Relevance feedback, explaining the user's query concept, can be used to bridge the semantic gap and improve the performance of CBIR systems. This paper proposes a learning method for relevance feedback, which develops logistic regression models to generalize the 2-class problem and provide an estimate of probability of class membership. To build the model, relevance feedback is used as the training data and the iteratively reweighted least squares method is applied to estimate the parameters of the regression curve and compute the maximum likelihood. After logistic regression models are fitted, discriminating features are selected by the measure of goodness of fit statistics. The weights of those discriminating features are determined based on their individual contributions to the maximum likelihood. The probability of class membership can therefore be obtained for each image of the database. Experimental results show that the proposed learning method can effectively improve the average precision from 41% to 63% through five iterations of relevance feedback rounds.

1. INTRODUCTION

In hospitals and medical institutes, a large number of medical images are produced in ever increasing quantities and used for diagnostics and therapy. Content-based image retrieval (CBIR) has been proposed to address the problem of image retrieval from medical image databases. Content-based image retrieval refers to the retrieval of images whose contents are similar to a query example, using information derived from the images themselves, rather than relying on accompanying text indices or external annotation. One of the main challenges in CBIR is the semantic gap between low-level features that can be extracted from the images and the descriptions that are meaningful to users. To bridge this gap, machine learning methods have to be incorporated for automatic association of such descriptions to the low-level features. To identify what the user is looking for in the current session, the user has to be included in the retrieval process. The process of finding the user's target is called a retrieval session, which can be divided into several rounds. In each round the user provides feedback regarding the search results by identifying images as either relevant or irrelevant ones (i.e. relevance feedback). Based on this relevance feedback, the system learns the common visual features of the images and returns improved results to the user.

Content-based image retrieval has been proposed by the medical community for inclusion into picture archiving and communication systems (PACS), which integrates imaging modalities and interfaces with hospital and departmental information systems in order to manage the storage and distribution of images to radiologists, physicians, specialists, clinics, and imaging centers [2]. Image searching in PACS is currently carried out according to the alphanumerical order of textual attributes of images. However, the information which users are interested in is the visual content of medical images rather than that residing in alphanumerical format. The content of images is a powerful and direct query which can be used to search for images containing similar content. In addition, computer-aided diagnosis and case-based reasoning create strong needs for CBIR technology.

The contributions of this study are to present a complete framework of content-based mammogram retrieval and, more importantly, propose an effective learning method for relevance feedback. The proposed content-based mammogram retrieval methods can be applied to efficiently retrieve those mammograms with similar pathological characteristics from distributed mammogram databases at hospitals and breast screening centers connected together through PACS. The remainder of this paper is organized as follows. An overview of the proposed content-based retrieval framework is described in Section 2. The proposed learning method for relevance feedback is developed in Section 3. Experimental setup and results

are presented in Sections 4 and 5. Section 6 presents the conclusions of this study.

2. OVERVIEW OF THE PROPOSED IMAGE RETRIEVAL FRAMEWORK

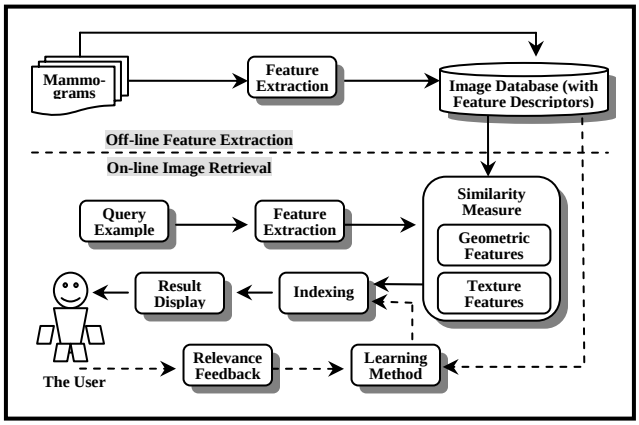

Figure 1. The proposed framework for image retrieval.

The proposed content-based retrieval framework as shown in Figure 1 can be divided into off-line feature extraction and on-line image retrieval. In off-line feature extraction, the contents of the images in the database are pre-processed, extracted and described with a feature vector, also called a descriptor. The feature vectors of the images constitute a feature dataset stored in the database. In on-line image retrieval, the user can submit a query example to the retrieval system to search for a desired image. The similarities between the feature vectors of the query example and those of the mammograms in the feature dataset are computed and ranked. It should be noted that feature vectors are divided into the geometric feature layer and the texture feature layer. The geometric feature layer is the first layer and a threshold value δ is set for selecting the prospective images. As the distance of similarity will be normalized into the range 0 to 1, the threshold is set in the range $0 \le \delta \le 1$. Only those prospective images are further considered for their similarity in the texture feature layer. Retrieval is conducted by applying an indexing scheme to provide an efficient way of searching the image database. Finally, the system ranks the similarity and returns the images that are most similar to the query example. This is called the initial search stage for a given query.

In abnormal mammograms, what doctors and radiologists are interested in are particular objects, such as calcium deposits and masses, because they are major signs of malignancy on mammograms. When comparing the similarity between two mammograms, characteristics of image content are taken into account based on their pathological importance. For this reason, similarity measures are divided into two different layers in the proposed system. The objective of the first layer is to sift out calcification mammograms from the whole image database and locate those mammograms with a similar size of calcium spots as the query example. The second layer considers the density of calcification spots and structure of breast tissue including density and distribution of fat, and directionality of breast muscle.

If the user is not satisfied with the initial search results, the user can conduct a relevance feedback stage. The user provides relevance feedback to the retrieval system in order to search further (following the dashed lines' arrows in Figure 1). To supply relevance feedback, the user simply identifies the positive image that is relevant to the query. The system subsequently analyzes the features of the user's feedback using a learning method and then returns refined results. This relevance feedback round can be repeated until the user is satisfied with the results or unwilling to offer any more feedbacks. The learning method is presented in Section 3.

3. LEARNING METHOD FOR RELEVANCE FEEDBACK

At the relevance feedback stage the task of the system is to learn user's relevance feedback, which are formed from user's subjective judgment on returned images. Common characteristics in relevant images reveal the user's search target and are what the user is interested in. To analyze the common characteristics and make a prediction, this study proposes a learning method shown in Figure 2.

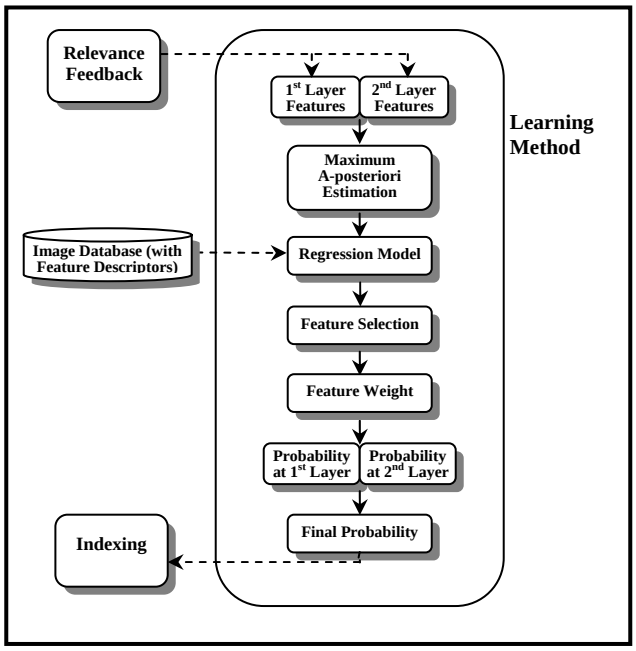

Figure 2. The proposed learning method for relevance feedback.

As described in Section 2, image features are divided into two different layers. The proposed method firstly collects relevance feedback as the training data, which is regarded as two different data sets. The first set includes the geometric features and the second set includes the texture features. The two sets of training data are then used to develop their individual logistic regression models, which are generalized linear models with binomial response and *logit* [3]. Next, the maximum *a-posteriori* method iteratively re-weighted least squares (IRLS) is applied to solve a least squares problem and estimate the parameters of the regression curve. When logistic regression models are developed, all database images are fit into the regressions. Discriminating features are then selected by the measure of goodness-of-fit statistics. The weights of the discriminating features are determined based on their individual contributions to the maximum likelihood. The probability of membership can be obtained for each image in the database. As two different feature layers are used to develop its individual regression model, each image can obtain two different probabilities for a different feature layer. The final probability that an image belongs to the relevant class will be obtained by multiplying these two probability rates. It is noted that, at the initial search stage, the similarity between any two images is measured by the mathematical distance of two points in the multi-dimensional system using a distance metric. At the relevance feedback stage, image similarity is completely based on the probability estimation. The detailed process is described in the following subsections.

3.1. Logistic Regression

Logistic regression is a mathematical modeling approach that can be used to describe the relationship of real-valued independent variables to a dichotomously dependent variable [4]. Suppose the response y of an image can take one of two possible values 0 and 1. $y = 1$ if the image is relevant to the query example; otherwise $y = 0$. Let $x = (x_1, x_2, ..., x_k)^T$ be the feature vector of an image. Since the output variable y only takes on values $\in \{0,1\}$ for the retrieval result, the logistic function can be used to represent $E(y \mid x)$ resulting in the range of $y \in \{0,1\}$. The logistic regression forms an s-shaped curve in which y approaches 1 as $x \to \infty$, or 0 as $x \to -\infty$. The *a-posteriori* probability of the class membership can be modeled via the linear function

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \qquad (1)$$

where $\beta = (\beta_0, \beta_1, ..., \beta_k)$ represents the weight vector, including the bias β_0. The conditional probability of label y given the feature vector x is expressed as

$$P(y = 1 \mid x, \beta) = \mu(x \mid \beta) = \frac{e^{\beta x}}{1 + e^{\beta x}} = \frac{1}{1 + e^{-\beta x}} \qquad (2)$$

$$P(y = 0 \mid x, \beta) = 1 - \mu(x \mid \beta) = \frac{e^{-\beta x}}{1 + e^{-\beta x}} \qquad (3)$$

The output of $f(x)$ is interpreted as an estimate of a probability $P(y = 1 \mid x, \beta)$ when equation (1) is transformed by the *logit* function.

$$logit\{\mu(x, \beta)\} = \ln\left[\frac{\mu(x, \beta)}{1 - \mu(x, \beta)}\right]$$

$$= \ln\left[\frac{\frac{1}{1 + e^{-\beta x}}}{\frac{e^{-\beta x}}{1 + e^{-\beta x}}}\right] = \ln\left[\frac{1}{e^{-\beta x}}\right] = \beta x \qquad (4)$$

The set of adjustable parameters β of the regression curve is key to developing the regression model. How to determine the parameters is described in the following section.

3.2. Maximum A-posteriori Estimation

Maximum *a-posteriori* estimation, which is maximizing the likelihood through the choice of the parameter estimates, is used to obtain estimates of the parameters for logistic regression. In this study, the iteratively re-weighted least squares (IRLS) algorithm is applied to solve a least squares problem in order to estimate the maximum *a-posteriori* parameters. The IRLS algorithm takes first and expected derivatives to obtain the score and information matrix and then develop a Fisher scoring procedure for maximizing the log-likelihood [4]. Suppose a logistic regression model is fit to a set of n samples $X = (x^1, y^1), \ldots, (x^n, y^n)$, which are randomly drawn from a binomial distribution. The conditional likelihood of a single observation can be expressed as

$$P(y^i \mid x^i, \beta) = \mu(x^i \mid \beta)^{y^i} (1 - \mu(x^i \mid \beta))^{1 - y^i} \qquad (5)$$

Then, the conditional likelihood of the whole dataset is written as

$$P(Y \mid X, \beta) = \prod_{i=1}^{n} \mu(x^i \mid \beta)^{y^i} (1 - \mu(x^i \mid \beta))^{1 - y^i} \qquad (6)$$

The conditional log-likelihood is

$$l(\beta \mid X, Y) = \sum_{i=1}^{n} (y^i \log \mu(x^i \mid \beta) + (1 - y^i) \log(1 - \mu(x^i \mid \beta)) \qquad (7)$$

The conditional log-likelihood equation can be solved by taking gradients

$$\frac{\partial l(\beta \mid X, Y)}{\partial \beta_j} = \sum_{i=1}^{n} y^i \left(\frac{1}{\mu(x^i \mid \beta)} \frac{\partial \mu(x_i \mid \beta)}{\partial \beta_j} - (1 - y^i) \frac{1}{(1 - \mu(x^i \mid \beta))} \frac{\partial \mu(x^i \mid \beta)}{\partial \beta_j} \right)$$

where $j \in [0, k]$. (8)

By plugging $\mu(x \mid \beta)$ of Equation (2) in Equation (8) and let the results equal to 0, a set of nonlinear equations can be obtained as

$$\sum_{i=1}^{n} x_k^i (y^i - \frac{1}{1+e^{-\beta x^i}}) = 0 \quad (9)$$

Newton's method is a general procedure for finding the roots of an equation $f(\theta) = 0$ based on the recursion [4].

$$\theta_{t+1} = \theta_t - \frac{f(\theta_t)}{f'(\theta_t)} \quad (10)$$

Although Newton's method is to find the minimum of a function *f*, the maximum of a function $f(\theta)$ is given when its derivative $f'(\theta) = 0$. Hence, plugging in $f'(\theta)$ for $f(\theta)$ above, it results in

$$\theta_{t+1} = \theta_t - \frac{f'(\theta_t)}{f''(\theta_t)} \quad (11)$$

The parameter β is a vector in logistic regression. The Newton-Raphson algorithm is introduced [4]

$$\beta_{t+1} = \beta_t - H^{-1} \nabla_\beta l(\beta_t \mid X, Y) \quad (12)$$

where $\nabla_\beta l(\beta_t \mid X, Y)$ represents the vector of partial derivatives of the log-likelihood equation, and $H_{ij} = \frac{\partial^2 l(\beta_t \mid X, Y)}{\partial \beta_i \beta_j}$ represents the Hessian matrix of second order derivatives. Then, Fisher scoring is applied to find the solution to the conditional log-likelihood equation. Taking the second derivative of the likelihood score equations gives us

$$\frac{\partial^2 l(\beta \mid X, Y)}{\partial \beta \beta^T} = -\sum_{i=1}^{n} x^i (x^i)^T \mu(x^i \mid \beta)(1 - \mu(x_i \mid \beta)) \quad (13)$$

The Newton-Raphson algorithm can be expressed using matrix notation for logistic regression. The $n \times n$ diagonal matrix is defined as

$$W = \begin{bmatrix} \mu(x_1 \mid \beta)(1-\mu(x_1 \mid \beta)) & 0 & \cdots & 0 \\ 0 & \mu(x_2 \mid \beta)(1-\mu(x_2 \mid \beta)) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mu(x_n \mid \beta)(1-\mu(x_n \mid \beta)) \end{bmatrix} \quad (14)$$

Let *Y* be an $n \times 1$ column vector of output values, and *X* be the design matrix of size $n \times (p+1)$ of input values, and *P* be the column vector of fitted probability values $\mu(x_i \mid \beta)$. The gradient of the log likelihood can be expressed in a matrix form as follows.

$$\frac{\partial l(\beta \mid X, Y)}{\partial \beta} = \sum_{i=1}^{n} x^i (y^i - \mu(x^i \mid \beta)) = X^T (Y - P) \quad (15)$$

The Hessian can be expressed as

$$\frac{\partial l(\beta \mid X, Y)}{\partial \beta_i \beta_i} = -X^T W X \quad (16)$$

The Newton-Raphson algorithm then becomes

$$\begin{aligned} \beta^{new} &= \beta^{old} + (X^T W X)^{-1} X^T (Y - P) \\ &= (X^T W X)^{-1} X^T W (X\beta^{old} + W^{-1}(Y - P)) \\ &= (X^T W X)^{-1} X^T W Z \\ &\text{where } Z \equiv X\beta^{old} + W^{-1}(Y - P). \end{aligned} \quad (17)$$

3.2. Feature Selection

Since a typical CBIR system only requires the user to submit images as query examples, no assumptions are made with regards to the characteristics of the content. To look for images with various characteristics, the system has to extract as many low-level features as possible from the images in the databases. Although discriminating features may be extracted, more redundant features may simultaneously undermine the retrieval performance if they are included into feature descriptors for further similarity measure. For this reason, this study proposes a feature selection process that automatically selects a subset of discriminating features at the relevance feedback stage. As the whole set of database images is fit to the regression model and the value of the maximum likelihood is estimated, this system then performs null hypothesis testing using a statistic measure of the goodness-of-fit. The statistic measure of the goodness-of-fit, also called a likelihood ratio test, compares a relatively more complex model, called a full model, to a simpler model, called a reduced model, to assess whether the simpler model fits the dataset significantly better. The likelihood ratio test begins with a measure of deviance between the full model that contains the observed features, and the reduced model that is the same as the full one except that the observed features are not included. The deviance can be compared to a chi-square distribution.

$$d_i = -2(\log L_f - \log L_r) \quad (18)$$

where L_f represents the maximum likelihood for the full model and L_r represents the maximum likelihood for the reduced model. If the full and the reduced models are both identical, the ratio of the full to the reduced model is one and its logarithm is zero, and therefore the deviance is zero. The likelihood test can assess whether the likelihood score is significantly improved as a specific feature is added into the model. To assess the importance of a specific feature x_i, the remaining features are used to account for it and that observed feature x_i is removed from the full model to obtain the reduced model. The null hypothesis is shown as follows:

$$H_0 : x_i = 0$$

$$H_1 : x_i \neq 0 \ \ (i = 1, 2, \ldots, k) \qquad (19)$$

H_0 refers to the reduced model with the observed feature = 0. H_1 refers to the full model. The predetermined α level of significance is set to 0.05 and the degrees of freedom (*df*) is 1. If the null hypothesis is accepted, it can be inferred that feature x_i does not have a significant effect on the maximum likelihood estimation because the deviance is too small. Otherwise, feature x_i has a significant effect on the complex model.

In addition to feature selection, this method further assesses the contribution from each discriminating feature and then determines the individual weighting based on their contributions. If feature x_i is assessed to have significant contributions on the maximum likelihood score, the weight w_i is computed via an exponential function

$$w_i = e^{-d_i}, \text{ where } d_i = -2(\log L_f - \log L_r) \qquad (20)$$

which results in $0 \leq w_i \leq 1$. If feature x_i makes a great contribution with an extremely large deviance value d_i, the weight w_i approaches one. On the contrary, the weight w_i approaches to zero.

3.3. Probability of Similarity

At the relevance feedback stage, the logistic regression can provide each image of the database with the *a-posteriori* probability P_c (i.e. probability at 1st layer in Figure 2.) of membership that is thought of as a relevant image. Similarly, each image obtains another probability P_t (i.e. probability at 2nd layer in Figure 2.) of membership that belongs to the relevant-image class. The final probability P_f is calculated as follows:

$$P_f = P_c P_t \qquad (21)$$

which represents each image's final score for the membership of the relevant-image class. Similar to the design in similarity measure described in Section 2, where a threshold value θ is set at the first feature layer, the threshold δ is also set at the 1st layer. As P_c represents a probability rate, the range of δ is $0 \leq \delta \leq 1$.

4. PERFORMANCE EVALUATION

The objective of this study is to apply the proposed learning method and framework to retrieve mammograms containing similar calcium deposits and structure of breast tissue

4.1. Mammogram Data Set

There are 750 images of size 200×200 pixels, each cropped from the Region Of Interest (ROI) of one mammogram. Among the 750 images, 250 of them contain calcification phenomenon while the other 500 do not. Calcification mammograms are classified into two different categories: micro-classifications and macro-classifications. Micro-calcifications, tiny calcium deposits less than 1/50 of an inch in size, are the most common mammographic sign of ductal carcinoma in situ (DCIS), which is a cancer that has not spread into neighboring breast tissue [1]. Macro-calcifications refer to larger, coarse calcium deposits.

4.2. Feature Extraction

To describe the mammogram contents, a total of 14 geometric and textural features were used in this study. Those features were separately used in two different layers.

Geometric Layer
It is observed that calcifications usually appear as spots which are the brightest areas when compared to the other breast tissues. Therefore, three operators with different sizes (3x3, 4x4, 5x5) are developed to detect the calcification spots, which creates three features [6]. Since the first layer is mainly used to verify whether an image has any calcium spots and the size of those calcium spots, geometric features are used to sift out similar mammograms.

Textural Layer
In an attempt to further discriminate the density of calcification spots and structure of breast tissue including density and distribution of fat, and directionality of breast muscle, the second layer makes use of textural features to analyze the image textures and compare the texture similarity between two images. The following features based on texture analysis were derived using the co-occurrence matrix: Contrast, Correlation, Inverse Difference Moment (also Homogeneity), Angular Second Moment (also Energy), Variance, Entropy, Different Variance, Different Entropy, Sum Average, Sum Variance, and Sum Entropy.

4.3. Relevance Feedback Process

In our experiments, 20 images were used as query examples to search for similar images. Among them, ten images contain micro-calcifications and the other ten contain macro-calcifications. Mammogram similarity is evaluated based on the existence of calcification, size of calcium spots, and distribution of calcifications. Relevance feedback is an interactive process and can have endless iterations. In CBIR, it is assumed that the users are unwilling to conduct too many relevance feedback rounds. Due to the assumption, the ability to rapidly learn the user's query concept is important for the system. Hence, it is worth observing whether the learning method can im-

prove the precision within a few iterations. Five iterations are conducted to learn the user's preference for each search session. In each round only relevant images are required to be identified as feedback, which will be accumulated throughout the whole session.

4.4. Experimental Results

This study investigates different values of the thresholds δ at the initial search stage and θ at the relevance feedback stage. In Table 1-4, iteration 0 indicates the precision at initial search stage before relevance feedback is involved. Iteration 1-5 indicates the precision at the relevance feedback stage. Three combinations (k1, k2, k3) of thresholds δ and θ are adopted to test the performance of our system. Combination k1 sets the most rigid threshold δ and θ. The results show that k1, k2, and k3 can obtain 62%, 64%, 62% on average precision. The results show that the average precision of macro-calcification can reach 61%; Micro-calcification can reach 65% by iteration 5.

Table 1. k1: Average precision for $\delta = 0.1$ & $\theta = 0.9$

Iteration	0	1	2	3	4	5
Macro.Precision	43%	51%	49%	52%	59%	58%
Micro.Precision	40%	41%	43%	48%	54%	67%
Ave.Precision	42%	46%	46%	50%	57%	62%

Table 2. k2: Average precision for $\delta = 0.2$ & $\theta = 0.8$

Iteration	0	1	2	3	4	5
Macro.Precision	41%	50%	49%	58%	61%	62%
Micro.Precision	41%	44%	51%	56%	53%	67%
Ave.Precision	41%	47%	50%	57%	57%	64%

Table 3. k3: Average precision for $\delta = 0.3$ & $\theta = 0.7$

Iteration	0	1	2	3	4	5
Macro.Precision	42%	53%	53%	59%	62%	62%
Micro.Precision	38%	41%	44%	51%	53%	62%
Ave.Precision	40%	47%	49%	55%	58%	62%

Table 4. Average precision rate obtained from k1, k2, & k3.

Iteration	0	1	2	3	4	5
Macro.Precision	42%	51%	50%	56%	61%	61%
Micro.Precision	40%	42%	46%	51%	54%	65%
Ave.Precision	41%	46%	48%	53%	57%	63%

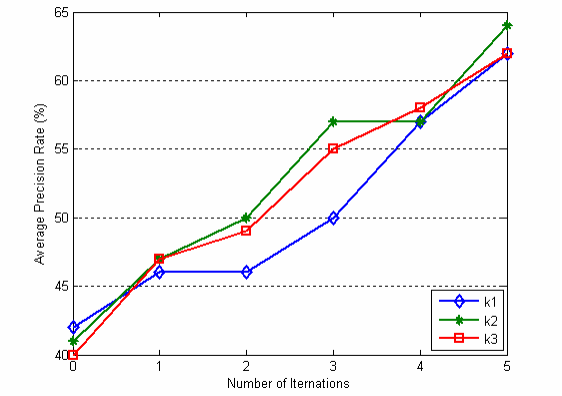

Figure 3. Average precision on different threshold values δ and θ. (Note: k1: $\delta = 0.1$ and $\theta = 0.9$; k2: $\delta = 0.2$ and $\theta = 0.8$; k3: $\delta = 0.3$ and $\theta = 0.7$)

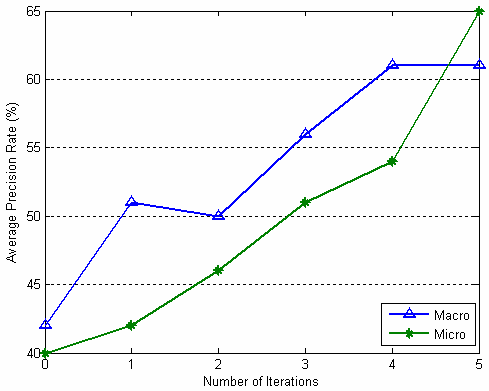

Figure 4. Comparison of average precision on macro-calcifications and micro-calcifications based on Table 4.

5. CONCLUSIONS

In this study, a learning method has been proposed to predict the probability of membership for content-based image retrieval. The proposed method develops logistic regression models, conducts features selection, and determines the weights of the discriminating features. The proposed method was evaluated using mammograms containing microcalcifications and macrocalcifications. The results show that the learning method can both learn two different types of calcification characteristics and improve the precision rate of image retrieval. Future work will explore other maximum likelihood estimation methods to rapidly obtain a smooth regression curve. A large variety of mammograms with pathological characteristics will be included into the image dataset to further investigate the learning and retrieval effectiveness.

7. REFERENCES

[1] El-Naqa, I., Yang, Y., Galatsanos, N. P., Nishikawa, R. M., and Wernick, M. N. "A similarity learning approach to content-based image retrieval: Application to digital mammography," *IEEE Transactions on Medical Imaging*, Vol. 23, No. 10, pp. 1233-1244, 2004.

[2] Huang, H. K., "PACS, image management, and imaging informatics," In D. Feng, W. C. Siu, & H. J. Zhang (Eds.), *Multimedia information retrieval and management: Technological fundamentals and applications*, Springer, New York, USA, pp. 347-365, 2003.

[3] Kleinbaum, D. G., *Logistic regression*, Springer-Verlag, New York, USA, 2002.

[4] Komarek, P., and Moor, A. W., "Making logistic regression a core data mining tool with TR-IRLS," *Proceedings of the fifth IEEE international conference on data mining*, Houston, USA, pp. 685-688, 2005.

[5] Wei, C.-H., and Li, C.-T., "Calcification descriptor and relevance feedback learning algorithms for content-based mammogram retrieval," *Proceedings of the international workshop on digital mammography 2006*, LNCS, Manchester, UK, 2006.