Statistical significance of A_z scores: Classification of masses in screening mammograms as benign or malignant based on high dimensional texture feature space

Gobert N. Lee and Murk J. Bottema
School of Informatics and Engineering
Flinders University
PO Box 2100, Adelaide SA 5001, Australia
and
Cooperative Research Centre for Sensor Signal and Information Processing
SPRI Building, Mawson Lakes Blvd, Mawson Lakes, SA 4095, Australia
glee@infoeng.flinders.edu.au
murkb@infoeng.flinders.edu.au

Abstract

In order to develop a method for classifying masses in digitised screening mammograms as benign or malignant, 260 image texture features were measured on 43 images of known malignant masses and 28 images of known benign masses. A genetic algorithm was used to select the optimal subset of k features based on A_z scores where k is a natural number. The leave-one-out A_z score for the optimal k features ranges from 0.80 to 0.95 for $k = 2, 3, ...12$. Since feature space reduction can result in optimistic estimates of classifier performance, the statistical significance of these scores were estimated by computing the empirical distribution of A_z scores in the context of the experimental parameters. For $k = 6, 7, 8$, the A_z scores were found to be significant at the $p = 0.05$ level.

1. Introduction

In the field of computer-assisted diagnosis, many authors have demonstrated that the texture of image intensity surfaces of screening mammograms provides information regarding the disease state of tissue [3, 4, 5]. Generally, these texture features are ones that are not seen by radiologists during visual inspection of the mammogram and are not based on models of the appearance of cancer in mammograms. Accordingly, the nature of texture features that are likely to provide positive predictive power is not well constrained. The result is that researchers are obliged to search far afield in order to discover optimal combinations of features. In addition, obtaining large numbers of training images on which to base the development of algorithms is not trivial and so a natural consequence is that studies comprise relatively small numbers of training images compared to the dimension of the feature space [2, 4].

Classification based on large number of features and a small training set can be optimistically biased. It can be shown that if the dimension of the feature space is greater than, or equal to, one less than the number of training images, then for any assignment of the training images into two groups, there exists a hyperplane which separates the two groups perfectly. Moreover, the hyperplane can be chosen so that the distance between an image in the feature space and the hyperplane (magnitude of the discriminant score) is the same for each training image.

One way to overcome this problem is to extract from the original feature space, a low-dimensional subspace that is realistic with respect to the size of the training data set. Selecting an arbitrary low-dimensional subspace defeats the purpose of considering many features, so the natural choice is a subspace that is optimal in some sense with respect to distinguishing between benign and malignant cases. However, the performance of the selected subspace in terms of classification is bound to be high. This is because in order to find the optimal subset, many different combinations of features are tested. It is expected that some of them will have a performance higher than average while others below average. From this collection, the feature combination that has the highest performance is selected, hence the performance will be high. The question is: is the performance of the optimal feature subspace selected greater than could be expected by chance?

The above question can be answered by performing a significance test. This will require knowledge of the distribution of the maximal performance scores obtained by repeating the selection process described above many times for data where there is no difference between the two groups. The distribution of these maximal A_z scores is not known and so was estimated using simulations.

Here we report on an experiment in which 260 texture features were measured on 71 training images. A genetic algorithm was used to select the k-dimensional feature subspace that is optimal with respect to A_z score for $k = 2, \ldots, 12$. The significance of the A_z score was measured by constructing empirical distributions of A_z scores for each k based on the full feature selection process.

2. Methods and materials

2.1 Data set

The data set comprises a total of 71 screening mammograms of which 43 contain malignant masses and 28 contain benign masses. The mammograms were obtained from the archives of *BreastScreenSA*, the South Australia branch of the National Screening Program in Australia. All malignant masses were biospy proven and the benign cases had a three years elapse time showing no sign of malignance. As the primary objective of the project is to assist diagnosis of clinically difficult cases, only the recall cases were included in the data set.

Electronic copies of the selected mammograms were acquired with a *Lumisys Lumiscan 150* laser digitiser. The resulting images have a spatial resolution of 50 μm and a depth resolution of 12 bits (4096 gray-level resolution). The images were reviewed and annotated by a radiologist experienced in mammography. Corresponding to the radiologist's annotation, regions of interest (ROIs) with a centering or near-centering mass were located. The size of each ROI is 1024×1024 pixels at full spatial resolution.

2.2 Texture measures

A total of 260 texture features were measured. These included 12 features based on image energy (see below), 8 based on gradients, and 240 based on co-occurrence matrices.

2.2.1 Textures Based on Co-occurrence Matrices

The co-occurrence matrix at distance d and direction θ is the array, P, where $P(i, j)$ is the joint probability that a pixel has image intensity value i and that the pixel at distance d in direction θ has value j. In addition to a choice of direction and distance, a co-occurrence matrix also requires a choice of quantisation of image intensity values. If the range of image intensity values is quantised to q bins, the co-occurrence matrix will be of size $q \times q$.

Co-occurrence matrices were constructed for distances $d = 11, 15, 21, 25, 31$, directions $\theta = 0, \pi/2$ and quantisation resolutions $q = 400, 100, 50$ for 40×40 half overlapping blocks in the straightened border region. The straightened border region, also called the rubber band straightened image, is an 80 pixel wide ring about the mass [4]. The directions $\theta = 0, \pi/2$ were chosen because they represent the directions perpendicular and parallel to the boundary of the mass. Radial structures near the mass boundary are known signatures of malignant masses. These 30 co-occurrence matrices were computed on two versions of the straightened border region. The first, called the polygon method, is found by connecting user defined points by line segments. The second, called the threshold method is found by finding a threshold for the ROI semi-automatically [2]. Hence a total of 60 co-occurrence matrices were constructed for every 40×40 block. The number of blocks varied from image to image depending on the size of the straightened border region, which, in turn varied according to the size of the mass.

For every co-occurrence matrix, the inverse distant moment (IDM) was computed according to the following formula.

$$IDM = \sum_{i=0}^{q-1} \sum_{j=0}^{q-1} \frac{1}{1 + (i - j)^2} P(i, j) \qquad (1)$$

For fixed values of d, θ, q, and choice of boundary method, the distribution of IDM values for all the 40×40 blocks in the straightened border region was recored. The first four moments of this distribution were recorded as features on which to base classification. Hence there were a total of 240 features based on co-occurrence matrices.

2.2.2 Intensity Gradient Features

The mass border was determined in each ROI using a polygon method. Background subtraction was performed. The geometric center of the polygon was used to define an 80 pixel wide annulus containing the border of the mass. The ROI was subsampled by a factor of 5 reducing the size from 1024×1024 to 204×204. The directional derivatives of the image intensity surface were computed in the directions both normal and tangential to the mass boundary. The first four moments of the distributions of the magnitudes of these directional derivates result in eight gradient features.

2.2.3 Local Image Energy Features

The local energy image y was computed from the image x by

$$y_{i,j} = \frac{1}{(2m+1)(2n+1)} \sum_{h=-n}^{n} \sum_{k=-m}^{m} x^2_{i-h,j-k}. \qquad (2)$$

For the region within the mass, the values $m = n = 12$ were used to produce an energy image restricted to the mass region. Two energy images were derived from the straightened border region. One with values $m = 3$ and $n = 10$, and the other with values $m = 10$ and $n = 3$. These choices were made to enhance features normal to the mass boundary in the first case, and tangential to the boundary in the second case.

For each of the three resulting energy images, the first four moments of the distribution of energy values were recorded resulting in a total of 12 image energy features on which to base classification and an over all total of 260 features from all three classes of features combined.

2.3 Genetic algorithm

A genetic algorithm [1] was used for feature subsets selection. The fitness criterion is based on the area under the receiver operating characteristic (ROC) curve, A_z, computed using the trapesoidal rule. (Technically, A_z referred to the area under a binormal ROC curve.) The genetic algorithm was initialised with a population of 1000 and is allowed to evolve over 500 generations. The mutation rate was set to 0.1. For each generation, the chromosomes with A_z score higher than the average A_z score of the current generation were retained in the parent pool. The remaining chromosomes were deleted from the population.

3. Classification results

The classification performance was evaluated using ROC methodology and the area under the ROC curve A_z was measured. As the optimal number of features k is not known a priori, classification using a range of feature number was performed. Figure 1 shows the training and the leave-one-out cross-validated A_z scores corresponding to $k = 2, 3, ...12$. The feature subsets correspond to the leave-one-out A_z scores are shown in Table 1. None of the optimal feature subsets include intensity gradient features, and therefore, these are not shown in Table 1.

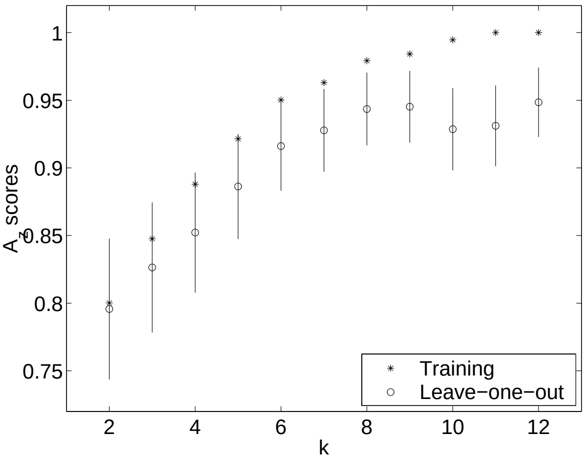

Figure 1. The classification results, both training and cross-validated A_z scores, are plotted against the number of features k. The cross-validated A_z scores are shown with error bars of one standard deviation.

4. Statistical significance estimation

In estimating the statistical significance of the classification results, the null hypothesis was that there is no difference between the two groups with respect to the k selected features. The null hypothesis implies that the observed classification results are no better than would be expected by chance. An empirical distribution of the maximal A_z scores based on the null hypothesis was simulated for each k. This was done by using a bootstrap method to generate 500 different 260 dimensional feature spaces with 71 data points and randomly assigning the data points to one group of 28 and the other of 43. For each of the 500 feature spaces, the genetic algorithm was used to search for the optimal k-dimensional subspace where $k = 3, 4, ...10$, and the associated optimal A_z score was recorded. Figure 2 shows the training maximal A_z distributions for $k = 3$ and $k = 10$. For $k = 4, 5, ...9$, the distributions were intermediate to the ones shown in the figure.

The statistical significance of the cross-validated A_z scores are the ones of interest since the cross-validated A_z scores provide a better (less biased) estimate of the classification performance. In order to estimate such statistical significance, the cross-validated A_z scores should be compared to the cross-validated A_z score distributions. Unfortunately, the cpu time needed to compute the cross-validated A_z score distributions was prohibitively large. Instead, the training A_z distributions were used. This gives a conservative estimate of the significance because A_z scores based on training data are positively biased when compared to the leave-one-out A_z scores. The statistical significance estimates of the leave-one-out A_z scores $k = 3, 4, ...10$ are shown in Figure 3. For $k = 6, 7, 8$ the A_z scores were found to be significantly larger than predicted by the null hypothesis at the $p = 0.05$ level.

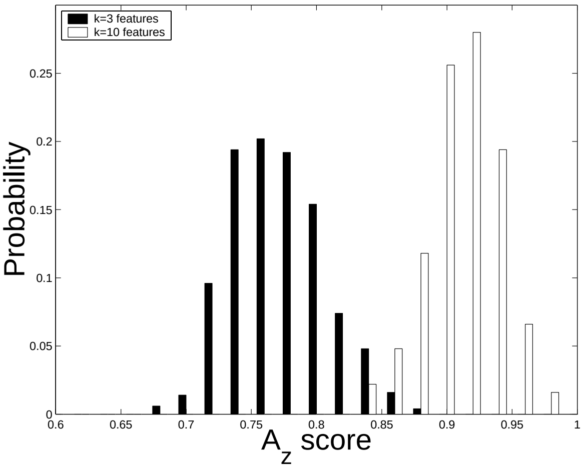

Figure 2. The empirical distributions of the training A_z score for $k = 3$ and 10. Each distribution consists of 500 data points.

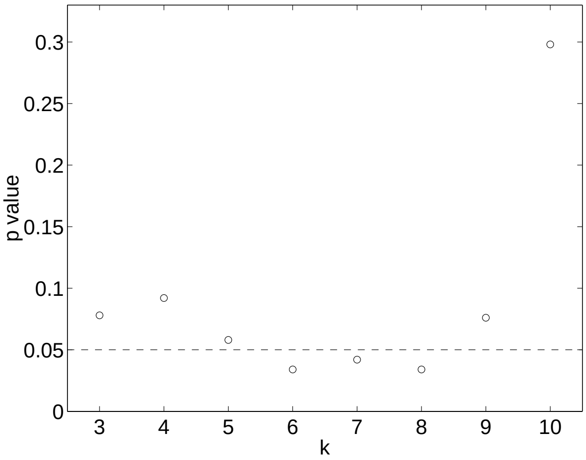

Figure 3. For each k, the p-value of the A_z score was computed by generating an empirical distribution of A_z scores (see text). For $k = 6, 7, 8$ the A_z values are significant at the $p = 0.05$ level.

5. Discussion and conclusion

Drawing conclusion regarding classification experiments comprising data sets that are small in size relative to the dimension of the feature space is always tenuous. Merely reporting the classification scores without incorporating the bias originating from the optimisation steps used to arrive at these scores does not provide sufficient information to judge the classification. Reporting confidence intervals for the classification performance score still does not acknowledge the bias of the method used to arrive at the score. This is borne out by the fact that the A_z score for some values of k were high but not significant. For example, for $k = 10$, $A_z = .9286$ but the p-value was near 0.3. By generating an empirical distribution of classifier performance values, it is possible to address the question of the significance of the classification performance measured experimentally.

References

[1] J. Holland. *Adaptation in natural and artificial systems.* Ann Arbor: The University of Michigan Press, 1975. Reprinted: Cambridge, Massachusetts: MIT Press, 1992/1994.

[2] G. Lee and M. Bottema. Classification of masses in screening mammograms as benign or malignant. In M. J. Yaffe, editor, *Proceedings of the 5th International Workshop on Digital Mammography, June 11-14, 2000, Toronto, Canada*, pages 259–263. Madison, Wisconsin: Medical Physics Publishing, 2001.

[3] I. Magnin, A. Bremond, F. Cluzeau, and O. C. Mammographic texture analysis - an evaluation of risk for developing breast cancer. *Optical Engineering*, 25(780–784), 1986.

[4] S. Sahiner, H.-P. Chan, N. Petrick, M. Helvie, and M. Goodsitt. Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. *Medical Physics*, 25(4):516–526, 1998.

[5] D. Thiele, T. Johnson, M. McCombs, and L. Bassett. Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes. *Medical Physics*, 23(4):549–555, 1996.

Table 1. Optimal feature subsets corresponding to the leave-one-out A_z scores in Figure 1. Table entries are the values of k for which that feature appeared in the optimal k feature subset. Q, d and θ are parameters of the co-occurrence matrices where Q is the gray-level scale quantisation, d is distance in pixels and θ is the direction measured anti-clockwise with 0 pointing vertically downward. M1, M2, M3 and M4 are the first 4 moments of the distribution of the inverse difference moment measured on co-occurrence matrices. Local images A,B and C are mass center region 25 $\times$ 25 pixels, straightened border regions 7 $\times$ 21 pixels and 21 $\times$ 7 pixels, respectively. m1, m2, m3 and m4 are the first four moments of the distribution of the energy values.

Co-occurrence matrix based features											Local image energy features				
			Border region by threshold method				Border region by polygon method								
Q	d	θ	M1	M2	M3	M4	M1	M2	M3	M4		m1	m2	m3	m4
400	31	0							3	4	A		7-12	9-11	
		$\pi/2$													
	25	0					8,10,12								
		$\pi/2$													
	21	0			2,4-12	3			2						
		$\pi/2$													
	15	0		12	11	8-10									
		$\pi/2$	6-7,12												
	11	0													
		$\pi/2$		4											
100	31	0					5-10,12				B				
		$\pi/2$													
	25	0					6-7,9								
		$\pi/2$	11		11										
	21	0				12									
		$\pi/2$													
	15	0				12									
		$\pi/2$	8-11	5											
	11	0													
		$\pi/2$	8-10	11	5-7,12										
50	31	0					11			11	C			12	
		$\pi/2$		3	4										
	25	0						11							
		$\pi/2$	8-10	5-7,12											
	21	0				12									
		$\pi/2$	10												
	15	0													
		$\pi/2$													
	11	0													
		$\pi/2$													