Dissimilarity Application in Digitized Mammographic Images Classification

U. Bottigli, B. Golosio, G.L. Masala, P. Oliva, S. Stumbo
Struttura Dipartimentale di Matematica e Fisica dell'Università di Sassari and Sezione INFN di Cagliari, Italy
Via Vienna 2, Sassari, 07100, Italy, fax: +39079229482; E-mail: giovanni.masala@ca.infn.it
and
D. Cascio, F. Fauci, R.Magro, G. Raso, M. Vasile
Dipartimento di Fisica e Tecnologie Relative dell'Università di Palermo and Sezione INFN di Catania, Italy
and
R. Bellotti, F. De Carlo, S.Tangaro
Dipartimento di Fisica dell' Università di Bari and Sezione INFN di Bari, Italy
and
I. De Mitri, G. De Nunzio, M. Quarta
Dipartimento di Fisica dell' Università di Lecce and Sezione INFN di Lecce, Italy
and
A. Preite Martinez , A. Tata
Dipartimento di Fisica dell'Università di Pisa and Sezione INFN di Pisa, Italy
and
P. Cerello, S.C.Cheran
Dipartimento d' Informatica dell' Università di Torino e Sezione INFN di Torino, Italy
and
E. Lopez Torres
CEADEN, Havana, Cuba

ABSTRACT

Purpose of this work is the development of an automatic classification system which could be useful for radiologists in the investigation of breast cancer. The software has been designed in the framework of the MAGIC-5 collaboration.
In the traditional way of learning from examples of objects the classifiers are built in a feature space. However, an alternative ways can be found by constructing decision rules on dissimilarity (distance) representations. In such a recognition process a new object is described by its distances to (a subset of) the training samples. The use of the dissimilarities is especially of interest when features are difficult to obtain or when they have a little discriminative power.
In the automatic classification system the suspicious regions with high probability to include a lesion are extracted from the image as regions of interest (ROIs). Each ROI is characterized by some features extracted from co-occurrence matrix containing spatial statistics information on ROI pixel grey tones. A dissimilarity representation of these features is made before the classification. A feed-forward neural network is employed to distinguish pathological records, from non-pathological ones by the new features. The results obtained in terms of sensitivity and specificity will be presented.

Keywords: Dissimilarity, Breast Cancer, Neural Network, Co-occurrence matrix, Computer Aided Detection.

1. INTRODUCTION

Breast cancer is reported as one of the first causes of women mortality [1] and an early diagnosis in asymptomatic women makes it possible the reduction of breast cancer mortality: in spite of a growing number of detected cancers, the death rate for this pathology decreased in the last 10 years [2], thanks also to early diagnosis, which has been made possible by screening programs [3].
MAGIC-5 (Medical Application on Grid Infrastructure Connection), a collaboration among italian physicists and radiologists, has built a large distributed database of digitized mammographic images and it is working on the development of CAD (Computer Aided Detection) tools for medical applications such as breast cancer detection through mammographic analysis, and lung cancer detection by Computed Tomography (CT) imaging modality. This collaboration has developed a system which, installed in an integrated station, can also be used for digitization, as archive and to perform statistical analysis. Furthermore this kind of station can represent also a very good system for mammographic educational programs. With a GRID configuration it would be possible for the clinicians tele- and co-working in new and innovative groupings. Using the whole database, several analysis can be performed by the MAGIC-5 tools.
The mammographic images (18x24 cm2, digitized by a CCD linear scanner with a 85 μm pitch and 4096 grey levels) are fully characterized: pathological ones have a consistent description which includes radiological diagnosis and histological data, while non pathological ones correspond to patients with a follow up of at least three years [4].The focus is on the automated analysis of masses, i.e. the search for rather 'large objects' in the image, usually characterized by peculiar shapes. The search is made using neural networks, with different algorithms of features extraction and with a different architectures.
We report in this work the results obtained in the classification of the region of interest (ROI) characterizing mass lesions. The novel approach is in the module of feature-extractor based on dissimilarity representation [5-8] of the features extracted from

co-occurrence matrix [9-10] containing second order spatial statistics information on ROI pixel grey levels.

2. METHODS

The CAD system here presented is an expert system based on three steps : a ROI-hunter, a features extractor module and a classifier based on neural network.

The ROI-hunter
The aim of this stage is to reduce the amount of data to process by searching for Regions Of Interest (ROIs), which are more likely to contain a mass. Only selected regions are retained for the next processing steps, rather than the whole mammogram, as shown in figure 1.

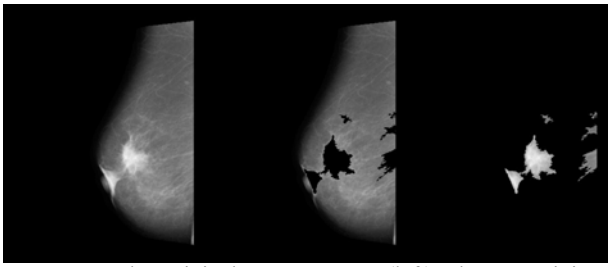

Figure 1: The original mammogram (left), the remaining image (middle), the selected patterns containing the ROIs (right).

The contour search is carried out by using the threshold operator. More generally this operator assigns the value I+ to the pixel with the intensity above a pre-fixed threshold, and the value I- to the pixel with the intensity less than the same threshold:

if $I_{x,y} \geq$ threshold $\Rightarrow$ $I'_{x,y} = I+$
if $I_{x,y} <$ threshold $\Rightarrow$ $I'_{x,y} = I-$

where $I_{x,y}$ is the intensity value for the pixel with coordinates (x,y).

An iterative procedure (*ROI Hunter*), based on the search of relative intensity maximum inside a square window, has been implemented to select the ROIs. In the literature the mass lesions typically vary in size from 2 - 40 mm in diameters [14]; in our case these two limits correspond to the square windows limit : A_{min} *(25x25)*, A_{max} *(501x501)*, in pixel. All the ROIs with area less than A_{min} are removed.

The steps of the algorithm are:

- starting from the right top corner of the mammogram, a raster scanning is performed to find the coordinates (x_o,y_o) of an intensity maximum I_m (the initial centre of the candidate lesion). Its value is accepted if it is also a relative maximum in a box A_{min} (25x25 pixels);
- an iso-intensity contour, including the relative maximum intensity pixel, is drawn at a threshold value $I_{th} = I_m / 2$; this contour defines a ROI with area A_R;
- the threshold Ith is dynamically changed by increasing/decreasing its value if the Area AR of the corresponding ROI is greater/smaller than the limit area Amax (501x501 pixels), until the difference between two consecutive thresholds is equal to one. At each step, the threshold is changed by an amount which is one half of the previous one.
- the ROI is removed and stored for a further analysis; the corresponding "hole" left in the mammogram is set to zero;
- go to the first step to find next (x_o,y_o) coordinates of a relative intensity maximum.

The number of ROIs detected from each image is related to the texture properties of the mammogram. All the ROIs extracted from negative images are tagged as negatives, while the ROIs from positive images can be labeled as true positive (TP) if they if they are overlapped with the contours of medical diagnosis. Otherwise as false positive (FP). A minimal rectangle, fully containing the ROI, is drawn with parallel sides with respect to the ones of the image to extract the features.

The features extractor
The module is composed by two steps :

- Feature extraction from co-occurrence matrix
- Dissimilarity representation

Feature extraction from co-occurrence matrix: In the first step, for each ROI we consider the minimal rectangular portion of the image which fully includes the ROI. The co-occurrence matrix is constructed from the image by estimating the pairwise statistics of pixel intensity, thus relying on the assumption that the texture content information of an image is contained in overall or average spatial relationship between pairs of pixel intensities [9].
Let us define the distance d between two pixels of the image as the minimum number of steps for going from one pixel to the other, where steps in the horizontal, vertical and diagonal directions are allowed. Two pixels at distances d and polar angle α are said to have a polar separation (d, α) [8].
Let G be the number of grey levels in the image (G = 2n for an n-bit image). For a given polar separation (d,α) a co-occurrence matrix M is a G x G matrix, which elements pij represent the fraction of pixels with grey levels i and j and polar separation (d, α) [9]. An example is represented in figure 2. In our work we considered only displacements d = 1 at quantized angles $\alpha = k\pi/4$, with k = 0,1,2,3.

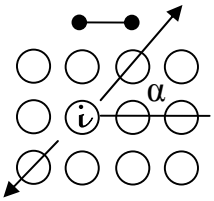

Figure 2 : polar separation (d, α)

Textural features can be derived from the co-occurrence matrix and used in texture classification in place of the single co-occurrence matrix elements. In ref. [15-16] 4 features are introduced, related to a textural property of the image such as

homogeneity, contrast, entropy and energy. The values of these features are sensitive to the choice of the direction α.
The usefulness of these parameters is to extract the informative content of matrix M in order to supply the features usable for the characterization of the texture: contrast supplies the indication of the more meaningful answer to the operator α.
Homogeneity and entropy supply the indication on the dominancy of the values on the main diagonal, on the base of the frequencies of the problem. Energy parameter supplies the information on the randomness of the spatial distribution.

The features used are in table 1:

contrast:	$\sum_{i,j}(i-j)^2 \cdot p(i,j)$
homogeneity:	$\sum_{i,j}\frac{p(i,j)}{1+\lvert i-j\rvert}$
entropy:	$-\sum_{i,j}\ln[p(i,j)]\cdot p(i,j)$
energy:	$\sum_{i,j}p(i,j)^2$

Table 1: Textural features used

So using 4 co-occurrence matrices ($\alpha = k\pi/4$, with k = 0,1,2,3) and 4 features for each matrix the record to be classified is composed by 16 features.

Dissimilarity representation: In the second step the dissimilarity representation is made. The representation based on dissimilarity [5-7] relations between objects is an alternative to the feature-based description. In general, dissimilarities are built directly on raw or pre-processed measurements, e.g. based on template matching. A dissimilarity value expresses [6] a magnitude of difference between two objects and becomes zero only when they are identical.
To construct a decision rule on dissimilarities [5], the interesting set T with n elements and the representation set R with r elements will be used. R consists of prototypes which are representatives of all involved classes. In the learning process, a classifier is built on the n x r dissimilarity matrix D(T,R), relating all training objects to all prototypes. The information on a set S of s new objects is provided in terms of their distances to R, i.e. as an s x r matrix D(S,R).
It is simply more important that the measure itself is discriminative for the classes than its strict metric properties. However, many traditional prototype optimization methods are not appropriate for non-metric dissimilarities, especially if no accompanying feature-based representation is available, as they often rely on the triangle inequality [12].
When only n samples [6] are available in an n-dimensional space, they are not sufficient for representing the real data distribution. It is known [12] that the feature-based classifier can perform poorly. Therefore, reduction of the dimensionality is important, also because of the computational aspect when the test sample is considered.

In our case the Euclidean distance [8] and a representative set R composed by r = 24 records with m = 16 features (characterizing the ROI) are chosen. The R set is composed by 12 healthy ROIs and 12 pathological ROIs extracted from several good images (with different tissue, type of mass lesions, projection, side, and other tips) which are a good database sampling.

Modelling the classes: A better characterization is made using 4 classes to distinguish masses. Therefore 5-classes are considered, where class 0 is the healthy one and classes 1,2,3 and 4 are various types of mass lesions as in the figure 2.

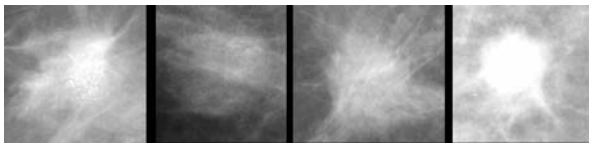

Figure 2: Some examples of masses included in the representative set extracted from the MAGIC-5 database. From left to right: speculated lesions, roundish lesions with regular, irregular, and blurred edge.

In the international literature [1-4],[13]-[14] the problem of the masses is normally dealt as a 2 class problem where the interesting discrimination is between healthy ROI and pathological ROI. The idea is to extend the classes to 5, starting from the presumption that four are the main typologies of masses distinguished from the radiologist. The interest is not to create an expert system able to classify between 5 types of opacities (including the healthy case) but to force the classifier to better distinguish the masses between them. In fact, the 4 typologies of masses are more varied between them (e.g.: shape - round with regular or irregular edges, opacity to star shape etc., as explained in chapter 2) that properly they cannot have the same positioning in the space of the information. To distinguish the masses between them it can be useful to better discriminate the healthy cases from the pathological cases. Therefore, it is not important if at the end the classifier gets confused between masses of various classes because the method is judged (good) if, altogether, the healthy ROI are separated from the pathological ones. This is a way to create data clustering based on the models of the classes. In this work, it was decided to use a model of 5 classes working with the new features representation with the dissimilarity discussed in the following paragraph (5.2). Preliminary studies involving 5 classes classification (without dissimilarity representation) do not supply substantial improvements of the classifiers performances.

Making dissimilarity: The dissimilarity representation and the reduction of the dimensionality is made by the following two steps:

→ Calculation of the distance for each record i of the interesting set T to each record k of the representation set R. Each record of T and R is a vector with m elements (number of features):

$T_i = (t_{i1},t_{i2},..,t_{im})$ $\quad i = 1,..,n$ (1)

with n defined as the number of records (ROIs) of the set T

$R_k = (r_{k1},r_{k2},..r_{km})$ $\quad k = 1,..,r$ (2)

with r = 24 defined as the number of records (ROIs) of the set R

$$d_{ik}{}^{j} = \sqrt{\sum_{m} (t_m - r_m)^2} \quad (3)$$

with m = 1,..,16, k = 1,..,r and j = 0,..,4 the class of the R set

→ For each record i of the set of interest, the class j of each record k of the R set is known to the expert system, while the classes of the T set are unknown.
For each record i of the interesting set T we can build the vector of the minimum distances from all records of R in the class j, so to obtain a features reduction :

$$d_i = (d_{min}{}^{0}, d_{min}{}^{1}, , d_{min}{}^{2}, d_{min}{}^{3}, d_{min}{}^{4}) \quad (4)$$

The classifier: After dissimilarity representation a multi-class problem is solved (5-classes). We make a study with a neural network classifier. The selected supervised classifier is a feed-forward neural network (FF-NN) trained with the back-propagation algorithm. It is used the gradient descent learning rule with "momentum", so as to quickly move along the direction of decreasing gradient, thus avoiding oscillations around secondary minima.

Own feed-forward neural network has 5 input, 7 hidden and 5 output neurons. The final output is 0 (healthy ROI) if the FF-NN answers class 0 and is 1 (pathological ROI) if the FF-NN answers with each other pathological classes (1,2,3,4).

The dataset extracted from the CALMA database [4] is shown in the table 2 and all results are validated with the k-folder (k = 5) cross-validation.

	Pathological sample (class 1, class 2, class 3, class 4)	Healthy sample (class 0)
Training set record 235	145 (42,34,44,25)	90
Test set record 238	147 (67,46,32,2)	93

Table 2: Dataset of ROIs extracted from CALMA database

3. RESULTS

Using sensitivity (percentage of pathologic ROIs correctly classified) and specificity (percentage of non pathologic ROIs correctly classified), the results obtained with this analysis are described in terms of the ROC (Receiver Operating Characteristic) curve [17-18], which shows the true positive fraction (sensitivity), as a function of the false positive fraction (1-specificity) obtained varying the threshold level of the ROI selection procedure. In this way, the ROC curve produced allows the radiologist to detect masses with predictable performance, so that he can set the desired true-positives fraction value and know the corresponding false-positives fraction value.

The best results of the automated masses analysis are about 91% for sensitivity and 67% for specificity. The ROC curve is shown in figure 3 and in table 3 the best ROC points are shown.

Sensitivity %	Specificity %
77.00	77.00
80.15	76.00
85.64	72.65
89.40	68.74
90.77	67.05
92.82	64.81
94.87	61.44
96.58	59.21

Table 3 : In the table are shown the best point in the ROC curve with dissimilarity application.

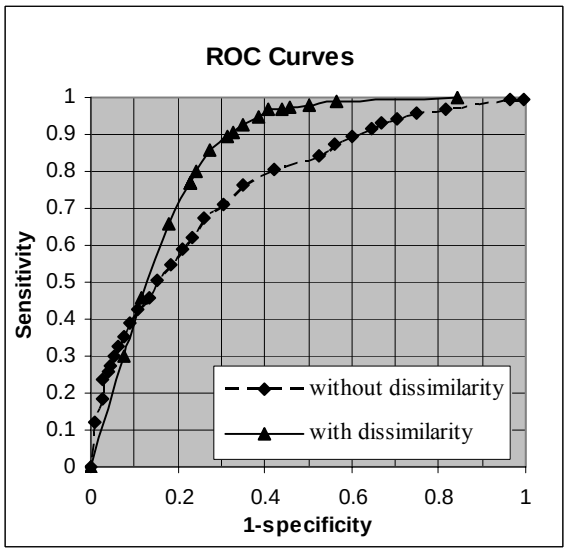

Figure 3 : ROC curve for the FF-NN

The overall performance is evaluated in term of the area under the ROC curve obtaining with dissimilarity application:

$$Az = (81 \pm 2.7)\,\%$$

The results of the same classifier optimized without dissimilarity representation supply an area under ROC curve :

$$Az = (76 \pm 2.8)\,\%$$

Therefore the application of the dissimilarity representation considerably increases the performances of the classifier.
Ulterior tests with reduction of the features to the point 2) through notion of median or medium distance in place of that minimum have turned out less convenient while the Euclidean distance of the point 1) has verified the triangular inequality. At the moment other notion of distance, beyond that Euclidean, has been tried with poor results.

4. CONCLUSION

In this paper an algorithm for masses classification has been presented. This algorithm is based on features extracted from co-occurrence matrix, containing second order spatial statistics information on ROI pixel grey tones and uses a dissimilarity representation. The new reduced features, in terms of minimum distances from a prototype set, are used for the discrimination between the two classes (pathological or healthy ROIs). The real interest for the radiologist are two class problems, therefore, the low accuracy of the classifier in 5 class problems is not important. Furthermore, the five classes division is made only to improve the difference between the four pathological classes then non pathological class by dissimilarity representation.

The discriminating performances of the algorithm was checked by means of a supervised neural network and the results have been presented in terms of ROC curve. The application of the dissimilarity representation considerably increases the performances of the classifier respect to the case without dissimilarity application.

The results are comparable than those obtained in other recent studies [4][11][13-14] verifying that the dissimilarity representation applied to the co-occurrence matrices provides a better ability to distinguish pathological ROIs from the healthy ones.

5. REFERENCES

[1] R.A. Smith, "Epidemiology of breast cancer", in **A categorical course in physics. Imaging considerations and medical physics responsibilities**, Madison, Winsconsin, Medical Physics Publishing, 1991.

[2] R. Peto, J. Boreham, M. Clarke, C. Davies., V. Beral, correspondence "UK and USA Breast cancer deaths down 25% in year 2000 at ages 20-69 years", **LANCET**, 355, (9217) , 2000, pp. 1822-1823.

[3] Blanks, **British Medical Journal** 321, 2000, pp. 655-659.

[4] R. Bellotti, F. De Carlo, G. Gargano, G. Maggipinto, S. Tangaro, M. Castellano, R. Massafra, D. Cascio, F. Fauci, R. Magro, G. Raso, A. Lauria, G. Forni, S. Bagnasco, P. Cerello, S. C. Cheran, E. Lopez Torres, U. Bottigli, G. L. Masala, P. Oliva, A. Retico, M. E. Fantacci, R. Cataldo, De Mitri I., G. De Nunzio, "A completely automated CAD system for mass detection in a large mammographic database" on **Medical Physics**, Vol. 33, No. 8 , Aug 2006, pp. 3066-3075.

[5] E. Pekalska, R. P. W. Duin "On Combining Dissimilarity Representations ", on **Multiple Classifier System** Second International Workshop, MCS 2001 Cambridge, UK, July 2001, pp 359-368.

[6] E. Pekalska, R. P. W. Duin "Classifiers for dissimilarity-based pattern recognition" 3nd International Conference on Pattern Recognition Barcelona September 2000 , vol 2 **Pattern Recognition and Neural Networks**, pp.12-16.

[7] E. Pekalska, R.P.W. Duin, R.P.W. and P.Paclik, "Prototype Selection for Dissimilarity-based Classifiers", **Pattern Recognition**, February 2006, vol. 39, no. 2, pp. 189-208.

[8] O. Duda, P. E. Hart, D. G. Stark, **Pattern Classification**, second edition, A Wiley-Interscience Publication John Wiley & Sons, 2001.

[9] R. M. Haralik, K. Shanmugam, I. Dinstein, "Textural Features for Image Classification" **IEEE Transactions on systems, Man and Cybernetics**, Vol. SMC-3, NO. 6, November 1973.

[10] R. W. Conners and C. A. Harlow. "A Theoretical Comparison of Texture Algorithm", **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2:204-222, 1980.

[11] F. Fauci, S. Bagnasco, R. Bellotti, D. Cascio, S. C. Cheran, F. De Carlo, G. De Nunzio, M. E. Fantacci, G. Forni, A. Lauria, E.Lopez Torres, R. Magro, G. L. Masala, P.Oliva, M. Quarta, G. Raso, A. Retico, S.Tangaro, "Mammogram Segmentation by Contour Searching and Massive Lesion Classification with Neural Network", **IEEE-Transactions on Nuclear Science (TNS)** Vol. 53, No. 4 ,August ,2006.

[12] M. Skurichina and R. P. Duin "Bagging for linear classifier", **Pattern Recognition**, 31(7); 909-930, 1998.

[13] A.H. Baydush, D.M. Catarious Jr, C .K .Abbey, C.E. Floyd, "Computer aided detection of masses in mammography using subregion Hotelling observers", **Medical Physics**,30, 2003, pp.1781-1787.

[14] G.D. Tourassi, R. Vargas-Voracek, D. M. Catarious Jr, C.E. Floyd Jr, "Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information", **Medical Physics**: 30(8) , 2003, pp. 2123-2130.

[15] D. H. Ballard, C. M. Brown, **Computer Vision**, Parentice Hall, 1982.

[16] J. Serra, **Image Analysis and mathematical morphology**, New York, NY, Academic Press,1983.

[17] J. A. Hanley, B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve", **Radiology**: 143,1982, pp. 29-36,.

[18] J.A. Hanley ,B. McNei, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases", **Radiology** 1983: 148; pp. 839-843.