IL NUOVO CIMENTO Vol. 28 C, N. 6 Novembre-Dicembre 2005
DOI 10.1393/ncc/i2005-10162-9

Classifiers trained on dissimilarity representation of medical pattern: A comparative study(*)

G. L. Masala(1)(2)(**), B. Golosio(1)(2), P. Oliva(1)(2), D. Cascio(3)(4)
F. Fauci(3)(4), S. Tangaro(5), M. Quarta(6), S. C. Cheran(7) and
E. Lopez Torres(8)

(1) *Struttura Dipartimentale di Matematica e Fisica dell'Università di Sassari - Sassari, Italy*
(2) *INFN, Sezione di Cagliari - Cagliari, Italy*
(3) *Dipartimento di Fisica e Tecnologie Relative dell'Università di Palermo - Palermo, Italy*
(4) *INFN, Sezione di Catania - Catania, Italy*
(5) *Dipartimento di Fisica dell' Università di Bari and INFN, Sezione di Bari - Bari, Italy*
(6) *Dipartimento di Fisica dell' Università di Lecce and INFN, Sezione di Lecce - Lecce, Italy*
(7) *Dipartimento d'Informatica dell' Università di Torino and INFN, Sezione di Torino Torino, Italy*
(8) *CEADEN - Havana, Cuba*

(ricevuto il 22 Luglio 2005; approvato il 4 Ottobre 2005; pubblicato online il 18 Gennaio 2006)

Summary. — In this paper we investigate the feasibility of some typical techniques of pattern recognition for the classification of medical examples. The learning of the classifiers is not made in the traditional features space but it can be made 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. 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 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 gray tones. A dissimilarity representation of these features is made before the classification. A Feed-Forward Neural Network (FF-NN), a K-Nearest Neighbour (K-NN) and a Linear Discriminant Analysis (LDA) are employed to distinguish pathological records from not-pathological ones by the new features. The results obtained in terms of sensitivity (percentage of pathological ROIs correctly classified) and specificity (percentage of healthy ROIs correctly classified) will be comparatively presented. The K-NN classifier gives slightly better results than FF-NN and LDA accuracy (percentage of cases correctly classified) on two-classes problem (pathologic or healthy patients).

PACS 87.57.Ra – Computer-aided diagnosis.
PACS 87.58.Mj – Digital imaging.
PACS 87.57.Nk – Image analysis.
PACS 87.59.Ek – Mammography.

(*) The authors of this paper have agreed to not receive the proofs for correction.
(**) E-mail: giovanni.masala@ca.infn.it

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 is working on the development of CAD tools for medical applications such as breast cancer detection through mammographic analysis. 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. Using the whole database, several analysis can be performed by the MAGIC-5 tools.

The mammographic images (18×24 cm^2, digitized by a CCD linear scanner with a 85 μm pitch and 4096 gray levels) are fully characterized: pathological ones have a consistent description which includes radiological diagnosis and histological data, while not pathological ones correspond to patients with a follow up of at least three years [4].The focus is on the automated analysis of massive lesions, *i.e.* the search for rather "large objects" in the image, usually characterized by peculiar shapes. The search is made using several classifiers of the pattern recognition, with the same algorithms of features extraction and with a different architecture.

We report in this work the results obtained in the classification of the regions of interest (ROI) characterizing massive lesions. The use of dissimilarities is especially of interest when features are difficult to obtain or when they have a little discriminative power. 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. We present also the best classifiers performance of K-NN, FF-NN and LDA.

2. – Methods

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

The ROI-hunter is the same described in ref. [11]:

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 opacity. Only selected regions are retained for the next processing steps, rather than the whole mammogram as shown in fig. 1.

The features extractor module is composed by two steps:

– features extraction from co-occurrence matrix;

– dissimilarity representation.

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].

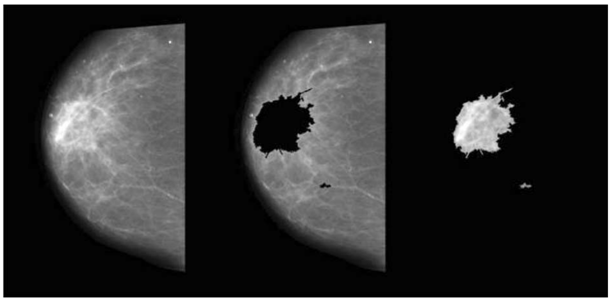


Fig. 1. – The original mammogram (left), the remaining image (middle), the selected patterns containing the ROIs (right).

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 ($G = 2^n$ for an n-bit image). For a given polar separation (d, α) a co-occurrence matrix M is a $G \times G$ matrix, which elements p_{ij} represent the fraction of pixels with grey levels i and j and polar separation (d, α) [9].

In our work we considered only displacements $d = 1$ at quantized angles $\alpha = k\pi/4$, with $k = 0, 1, 2, 3$.

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. [12, 13] 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 features used are in table I.

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.

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.

To construct a decision rule on dissimilarities [5, 6], 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

Table I.

Features used	
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	$\sum_{i,j} p(i,j)^2$

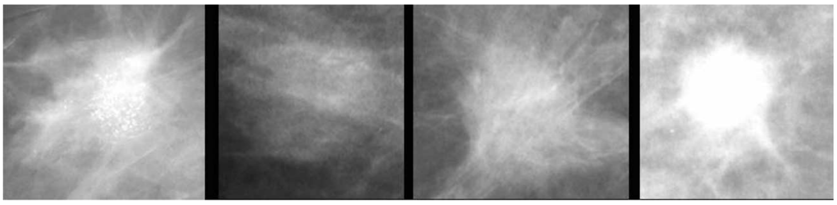


Fig. 2. – Some examples of opacities lesions included in the representative set extracted from the MAGIC-5 database. From left to right: speculated lesions, roundish lesions with regular, irregular, and blurred edge.

built on the $n \times r$ 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 $s \times r$ matrix $D(S,R)$. 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 massive lesions, projection, side, and other tips) which are a good database sampling.

A better characterization is made using 4 classes to distinguish massive lesions. Therefore 5-classes are considered (5-classes problem), where class 0 is the healthy one and classes 1, 2, 3 and 4 are various types of opacities as in fig. 2.

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):

$T_i = (t_{i1}, t_{i2}, \ldots, t_{im})$, $i = 1, \ldots, n$, with n defined as the number of records (ROIs) of the set T;

$R_k = (r_{k1}, r_{k2}, \ldots r_{km})$, $k = 1, \ldots, r$, with $r = 24$ defined as the number of records (ROIs) of the set R;

$d_{ik}^{j} = \sqrt{\sum\limits_{m} (t_m - r_m)^2}$, with $m = 1, \ldots, 16$, $k = 1, \ldots, r$ and $j = 0, \ldots, 4$ 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: $d_i = (d_{\min}^0, d_{\min}^1, d_{\min}^2, d_{\min}^3, d_{\min}^4)$.

After dissimilarity representation a multi-class problem is solved (5-classes) by the third step of the *classifiers*.

We make a comparative study of the following classifiers:

– A K–Nearest-Neighbors (K-NN) classifier. For this type of deterministic classifier, it is necessary to have a training set which is not too small, and a good discriminating distance. KNN performs well in multi-class simultaneous problem solving. There exists an optimal choice for the value of the parameter K, which brings to the best performance of the classifier. This value of K is often approximately close to $N^{1/2}$ [14]; in this work is $K = 9$.

– A Feed-Forward Neural Network (FF-NN) with 5 input, 7 hidden and 5 output neurons has been used. The selected FF-NN is a feed-forward back-propagation supervised network trained with gradient descent learning rule with "momentum", so as to quickly move along the direction of decreasing gradient, thus avoiding oscillations around secondary minima [8].

– A Linear Discriminant Analysis (LDA) is a classical statistical approach for classifying samples of unknown classes, based on training samples with known classes. LDA is the linear discriminant that maps the samples with known class from the n-dimensional (for us $n = 5$, *i.e.* number of input) space to the plane, in such a way that the ratio of the between-group variance and the within-group variance is maximized [8].

The final output for each classifier is 0 (healthy ROI) if the answer is class 0 and is 1 (pathological ROI) if the answer is with each other pathological class (1,2,3,4). The dataset extracted from the CALMA database [4] is shown in table II below and all results are validated with the k-folder ($k = 5$) cross-validation.

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 [15, 16], 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 massive lesions with predictable performance, so that he can set the desired true-positives fraction value and know the corresponding false-positives fraction value. The ROC curve is shown in fig. 3. In fig. 4 the classifiers performance in terms of specificity, sensitivity and accuracy in accuracy in five-classes problem is shown.

The overall performance is evaluated in terms of the area under the ROC curve obtaining for each classifier table III.

4. – Analysis and conclusion

In this paper an algorithm for massive lesion classification has been presented. 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

Table II.

Dataset for neural network

	Pathological sample Tot (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

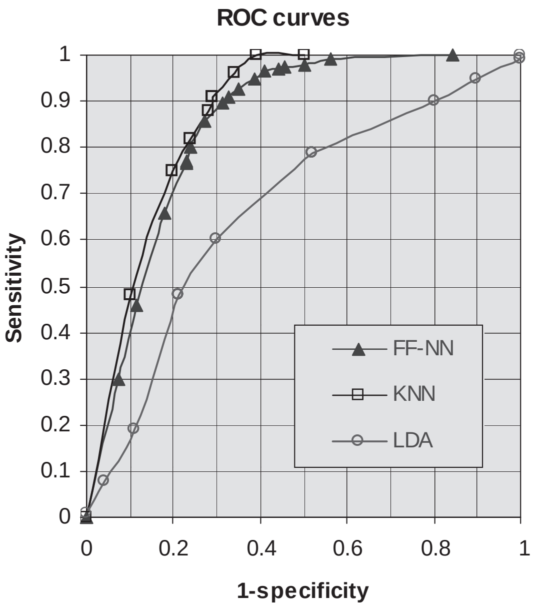
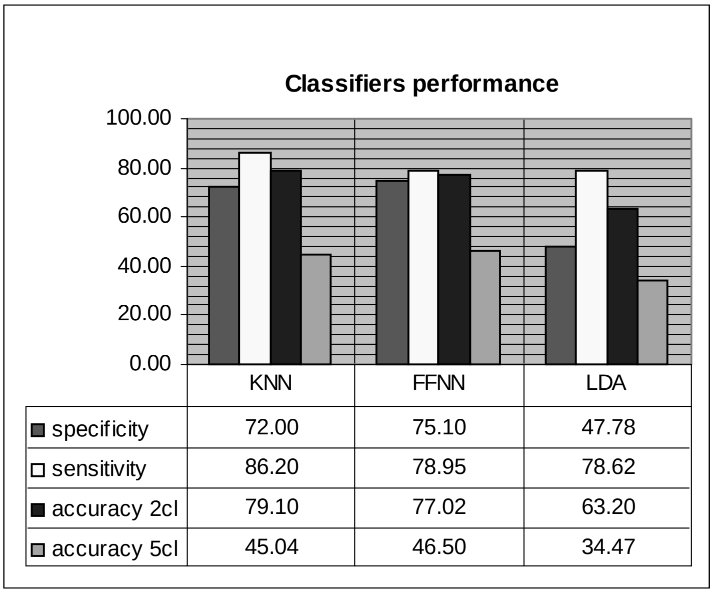


Fig. 3. – ROC curves for the classifiers MLP, LDA, KNN on the same representation of dataset.



	KNN	FFNN	LDA
■ specificity	72.00	75.10	47.78
□ sensitivity	86.20	78.95	78.62
■ accuracy 2cl	79.10	77.02	63.20
▨ accuracy 5cl	45.04	46.50	34.47

Fig. 4. – Classifiers performance in terms of specificity, sensitivity and accuracy in two-classes problem and accuracy in five-classes problem. The threshold level of the ROI selection procedure is zero for all classifiers.

Table III.

Classifiers	Area under the ROC curve
FF-NN	$A_z = (80.60 \pm 2.70)\%$
K-NN	$A_z = (86.56 \pm 2.69)\%$
LDA	$A_z = (65.76 \pm 3.06)\%$

discriminating performances of the algorithm were checked by a linear method as LDA, a statistics method as K-NN and a non-algorithms method as FF-NN. The best results in terms of area under the ROC curve and sensitivity are better for K-NN than the other classifiers. The low results obtained by LDA indicate that linear methods are not suitable for this medical problem.

The real interest for radiologist is for two classes problem so the low accuracy of the classifier in five-classes problem 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 results are comparable or better than those obtained in other recent studies [11, 17-19] verifying that the dissimilarity representation applied to the co-occurrence matrices provides a better ability to distinguish pathological ROIs from the healthy ones.

REFERENCES

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

[2] Peto R., Boreham J., Clarke M., Davies C. and Beral V., correspondence *LANCET*, **355** (2000) (9217) 1822.

[3] Bird R., Wallace T. and Yankaskas B., *Radiology*, **184** (1992) 613.

[4] Bottigli U., Delogu P., Fantacci M. E., Fauci F., Golosio B., Lauria A., Palmiero R., Raso G., Stumbo S. and Tangaro S., *Int. Soc. Opt. Eng. (SPIE)*, **4684** (2002) 1301.

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

[6] Pekalska E. and Duin R. P. W., *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] Duin R. P., *Prototype selection for dissimilarity-based classifier, Proceedings at WIRN 2004, Perugia 14-17 September 2004*, invited talk ISBN 88-89422-09-2.

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

[9] Haralik R. M., Shanmugam K. and Dinstein I., *Man. and Cybernetics*, Vol. SMC-3, No. 6, November 1973.

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

[11] Fauci F., Bagnasco S., Bellotti R., Cascio D., Cheran S. C., De Carlo F., De Nunzio G., Fantacci M. E., Forni G., Lauria A., Lopez Torres E., Magro R., Masala G. L., Oliva P., Quarta M., Raso G., Retico A. and Tangaro S., *Mammogram Segmentation by Contour Searching and Massive Lesion Classification with Neural Network, Proc. IEEE Medical Imaging Conference*, M2-373/1-5, *October 16-22, 2004, Rome, Italy*; in press.

[12] Ballard D. H. and Brown C. M., *Computer Vision* (Prentice Hall) 1982.

[13] Serra J., *Image Analysis and Mathematical Morphology* (Academic Press, New York, NY) 1983.

[14] Fukunaga K., *Introduction to Statistical Pattern Recognition*, second edition (Academic Press, Boston, MA) 1990.

[15] Hanley J. A. and McNeil B., *Radiology*, **143** (1982) 29.

[16] Hanley J. A. and McNeil B., *Radiology*, **148** (1983) 839.

[17] Skurichina M. and Duin R. P., *Pattern Recognition*, **31** (1998) 909.

[18] Baydush A. H., Catarious D. M. Jr, Abbey C. K. and Floyd C. E., *Med. Phys.*, **30** (2003) 1781.

[19] Tourassi G. D., Vargas-Voracek R., Catarious D. M. Jr and Floyd C. E. Jr, *Med. Phys.*, **30** (2003) 2123.