International Journal of Advances in Computer Science & Its Applications – IJCSIA
Volume 4: Issue 3 [ISSN 2250-3765]
Publication Date : 30 September,2014

A Simple, but Thorough Experimentation for Breast CE-MRI Classification

Carmela Luongo, Franco Alberto Cardillo, Giuseppe Amato

***Abstract*— We present the results of an experimentation with dynamic features for breast cancer detection in Contrast-Enhanced Magnetic Resonance of the female breast. In order to understand how good the various features are we built a dataset from real dataset with proven histological diagnosis. We compared human-readable features, commonly used in the clinical practice, with a non-linear artificial neural network trained with a double *k-fold* cross validation. The results show that the ANN reaches very good results when two specific dynamic features are used. The particular validation procedure used in this experimentation allows us to better understand the discriminative power of the various approaches and move toward a better classifier that might be used in the clinical environment. Breast cancer, in fact, is the second form of diagnosed cancer among women living in the western countries and any improvement in its diagnosis would lead to a lower mortality rate.**

Keywords—*medical image analysis, breast cancer, contrast-enhanced magnetic resonance imaging.*

I. Introduction

Breast cancer is the second form of diagnosed cancer and the second cause of cancer death among women living in the western world. Contrast-Enhanced Magnetic Resonance Imaging (CE-MRI) is an imaging modality which plays an important role in the diagnosis of breast cancer. In a Contrast-Enhanced MRI (CE-MRI), several series of images are acquired, each series containing 2-D slices of the entire volume containing the breast under examination. A first series is acquired without contrast agent (called the pre-contrast series), then several series are acquired after the injection (post-contrast series) of the contrast agent, whose diffusion in the tissues will cause some areas in the post-contrast images to appear 'brighter' than in the corresponding pre-contrast image. The diagnosis based on CE-MRI should take into account both morphological and dynamic features of the enhancing regions. In this experimentation we study the dynamic features, that captures the kinetic behaviour of the contrast agent diffusion in a small region, usually few voxels (i.e. volume element, it stands for pixel) wide, normally called Region Of Interest (ROI).

In this paper we present an experimentation about the discriminative capability of several dynamic features. Their values are used to train simple threshold-based classifiers and a multilayer perceptron. The rest of the paper is described as follows. In the next section we describe the CE-MRI examinations, the dynamic criteria, and the learning models used in the presented experimentation. In section III we summarize the results of the experimentation. In section IV we discuss the relevance of the dataset built during our study and the importance of the validation methods used in the experimental work.

II. Materials and Methods

All the CE-MRI datasets were acquired on a General Electric 1.5T Signa Contour scanner using 3D fast spoiled gradient echo sequences (FSPGR) with 12.7 ms repetition time, 2.5 echo time and 30° flip angle. Each study has six series acquired along the coronal plane. Besides the pre-contrast one, five series are acquired at zero, two, four, six, eight minutes after contrast agent injection. The number of images per series depends on the dimension of the acquired volume. The resonance signal is stored in DICOM files using nine bits per voxel. The most common dynamic feature used for breast CE-MRI image classification is known as relative enhancement $RE^{\%}$ and is computed as follows:

$$RE_k^{\%} = \frac{I_k(R) - I_0(R)}{I_0(R)}$$

where k is the post-contrast series index, I_k^j is the *j*-th image in a series (the index j has been omitted for the sake of clarity), I_o corresponds to the pre-contrast series, and $I_k(R)$ denotes the mean value of image I in series k over the region of interest (ROI) R. R is typically composed by only a few pixels, in our studies it corresponds to a 3×3 region. Since there are five post- contrast series, the relative enhancement values correspond to five points of a curve providing information about the diffusion of the contrast agent over time in a given ROI. When several curves extracted from malignant and benign ROIs are plotted, they

Carmela Luongo
Istituto Nazionale di Fisica Nucleare
Italy

Franco Alberto Cardillo, Giuseppe Amato
Istituto di Scienza e Tecnologie dell'Informazione - CNR
Italy

tend to form three basic clusters, as shown in **Figure 1**. The clusters can be described as follows:

Type I: a kinetic behaviour with a persistent uptake is considered a sign of benignity;

Type II: the classification of a kinetic behaviour with a plateau phase is uncertain;

Type III: a kinetic behaviour with a strong uptake followed by a washout is considered a sign of malignancy;

However, patterns extracted from the images present a kinetic behaviour quite different from the ideal curves in **Figure 1**. Real benign and malignant patterns overlap significantly. The other dynamic criteria investigated in the presented experimentation are described in the following.

Early increase. This criterion corresponds to the relative enhancement values 80 seconds after the injection of the contrast agent. Since the first post-contrast images are acquired zero and two minutes after the injection, this value is obtained by the available values with a simple linear interpolation. The value of 80 seconds corresponds to the time threshold used by many research and clinical groups, including the one working in the Pisa hospital, where the CE-MRI examinations used in this study have been acquired.

Washout Ratio [6]. This criterion tries to capture the washout rate in the late post-contrast phase and is defined as follows: $W_{peak-k} = \frac{I_{peak}(R) - I_k(R)}{I_{peak}(R)} \cdot 100$ where *peak* is the series index where (i.e., the minute when) the ROI *R* shows its maximum values. It represents basically the percentage in signal decrease between the series *k* and the series where ROI *R* reaches its maximum value.

Initial Slope [1, 6]. This criterion tries to capture the ratio between the maximum in relative enhancement and the time to that maximum. It is defined as follows: $Slope_i(R) = \frac{RE_{peak}(R)}{T_{peak}(R)}$ where *RE* is the relative enhancement value and *peak* is the index where the maximum *RE* is reached in ROI *R*.

Curvature at peak of enhancement [2]. This criterion tries to distinguish curves with the same peak of enhancement (same value at the same time) by studying the curvature of the relative enhancement curve when it reaches its maximum value. Its exact mathematical expression is quite complex, but can be roughly approximated and explained as: $k(T_{peak}) \approx A \cdot \alpha \cdot \beta$, where A is the upper limit of the signal intensity, α is the rate of signal increase, β is the rate of signal decrease during washout.

The experimental work presented in this paper is part of a larger research project that developed a Computer-Aided Diagnosis tool for the semi-automatic inspection of the CE-MRI studies. The tool includes algorithms for image filtering, segmentation, and registration. However, the image values used in this work correspond to the original values stored in the DICOM files by the MR scanner. Each slice, i.e. the two-dimensional image belonging to a series, is 256×256 pixels in size, each pixel corresponding to a volume element 1.5×1.5×3mm^3. The images are acquired along the coronal plane of acquisition and there are no gaps between slices.

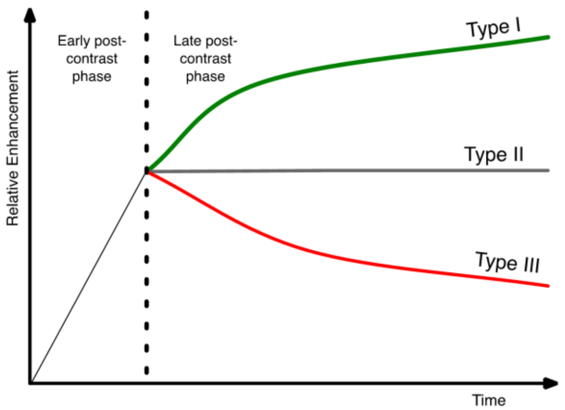

Figure 1. Typical (ideal) relative enhancement curves

In this study we experimented several dynamic features and studied how effective they are in discriminating benign and malignant lesions in the dataset we prepared. In order to build the dataset used in this experimentation, we used 40 CE-MRI examinations with a definite diagnosis proven by an histological analysis. 20 examinations contained benign lesions, the remaining 20 contained malignant lesions. We used only the two labels 'benign' and 'malignant' without adopting a finer classification with the different types of cancer since our goal was to simply find a reliable algorithm able to suggest regions worth a deeper inspection by a human operator. Among the 40 selected examinations, there are two false negative cases, i.e., examinations considered benign by the radiologists, but proven to be malignant by the histological analysis.

For each selected examination, an initial, large set of curves was extracted by the most enhancing areas from the volume regions indicated by the radiological and histological documentation. The extraction was performed automatically or semi-automatically by segmenting the images using anisotropic filtering, mathematical morphology, the watershed transform, and region growing methods. From such large sets of curves, 30 curves per examination were manually selected. The manual selection filtered out the curves presenting low or chaotic enhancement (usually extracted from veins). In fact, since the initial set of curves is extracted from regions segmented starting from voxels presenting a strong enhancement, it includes many curves that have no diagnostic value. It must be taken into account that not all of the curves extracted from a cancer region present strong enhancement.

The final dataset is thus composed by 30×40=1200 curves, half extracted from benign lesions and half from malignant lesions. This dataset has been used to train the classifiers, presented in the next section, in a supervised learning framework. In this context, a learning algorithm infers the classification rules by observing the input data associate to the correct label. In this case the labels are just two {*benign, malignant}*.

Two different types of classifiers have been used in this study. The first one is a basic threshold-based binary classifier used for the quantitative features. Such classifiers assign a final label in by comparing the input value to a threshold value. In this case, the learning step aims only at computing the best threshold able to separate the patterns in the two classes.

The second type of classifiers is the non-linear multilayer feedforward perceptron, an artificial neural network (ANN) based on the backpropagation learning algorithm [5]. This classifier has been used to study the qualitative feature “type of the curve” (see Fig. 1). A basic multilayer perceptron is composed by three layers of nodes: the input layer that receives the values to be processed and propagates their values to the second layer, the hidden layer. This layer basically constructs a representation of the values received during the training by the input layer, trying to learn the regularities characterizing the input dataset (it performs a kind of data compression). Each node sends its output to all the nodes in the subsequent layers (if any) and each node receives its input from all the nodes in the previous layer. During the training, the ANN observes input patterns labeled with the correct response and adjust its parameters in order to match the correct answer. The signal moves from the input layer to the output layer, where the output of then nodes is compared to the correct label. Based on this comparison, an error signal is sent back to the hidden and input layers that adapt their response in order to reduce the error as measured into the output layer.

A multilayer perceptron requires input values normalized in the or ranges, depending on the activation function of the nodes.

During the normalization of the input dataset, vectors differing only for a multiplicative constant are mapped onto the same normalized vector. The output of threshold-based classifiers, as well as of other classifiers that process the input values without normalization, needs to be used in combination with the ANN output in situations where the normalization step produces a loss of information.

III. Results

We experimented several configurations of the ANN classifiers, with different numbers of layers and using all or just a subset of the dynamic features described in the previous section. The experimentations involved all of the learning hyperparameters of the neural network: activation function (logistic or hyperbolic tangent), number of hidden nodes and training epochs, learning rate, and learning momentum.

By inspecting the histogram of the features described in the previous paragraphs, it is clear that only the criteria ``Washout ratio" and ``Relative Enhancement" can provide reliable information about the classification of the enhancement values. Threshold-based classifiers have been built for the criteria ``Washout ratio" and ``Initial Slope" and their performance tested when used in combination with the ANN.

The threshold based classifiers have been trained using a *k*-fold cross-validation with *k=10*. The *k*-fold cross-validation consists in partitioning the original dataset into folders and performing training steps. In each iteration, one of the subsets is selected as test set and the other *k - 1* are used as training set, thus using in each iteration a different training and test set. Therefore, the dataset has been partitioned into 10 subsets, each containing 120 input patterns balanced between the two classes. The performance of the threshold-based classifier is estimated by computing the mean of the performances obtained by the classifiers built by the *k*-fold cross-validation. The washout-ratio reaches *88.83%* mean accuracy, *88.16%* mean sensitivity, and a *90.5%* mean specificity. This result is used as a base reference to understand how better the ANN is at classifying vectors containing the relative-enhancement and the washout ratio values.

The multilayer perceptron has been experimented on several different datasets composed by all the dynamic features previously described or only by a subset of the same feature set. We cannot report all of the experimental results and we limit our description to the training method and the subset of features which showed the best performance.

First of all, in order to select the best hyperparameters of the neural network, we performed an initial test using the double *k*-fold cross-validation [4, 3]. The procedure consists in two nested cross-validations. In the external *k*-fold cross-validation, the dataset is partitioned into *k* balanced sets. The prediction error of the model on the test set is estimated as the mean of the error computed on each of the test sets (folders) in the cross validation iterations. The internal *l*-fold cross-validation is used for model selection and, in particular, to decide when the early stopping criterion is to be applied before the learning model overfits the input data. In the double *k*-fold cross-validation, the final results are computed for *lk* trained algorithms and can be considered a reliable estimate of the final prediction error. In our case we performed an external cross-validation with *10* folders and an internal one with 9 folders, thus providing a total of *90* trained neural networks.

Once computed the best values of the hyperparameters, a new neural network was trained with those values using a *10*-fold cross-validation. The *10* classifiers produced in this last training step compose a committee that is used in the

Computer Aided Diagnosis tool for classifying new CE-MRI of the breast.

The double *k*-fold cross-validation allows the single data to be used in all of the phases of the training process. The external *k*-fold cross-validation is used as described in the previous paragraph for the threshold-based classifier. The *k-1* folders used in the internal cross-validation are called calibration set. In our experimentation we used a *9*-fold cross-validation in the internal process. From these nine sets, eight are used for training and one for validating the training process. In our case we used that set for studying the training error of the neural network and stop the learning as soon as the training error started rising, meaning that the neural network was over-fitting the training data. All of the models were validated using the three classical measures: sensitivity, specificity, and accuracy. The model selection performed in the internal cross-validation selected a multilayer perceptron with the logistic activation function, a *0.1* learning rate, a *0.8* momentum, *30000* epochs, and early stopping set to . The ANN classification of relative-enhancement curves reaches a *91.5%* mean accuracy, 88.17% sensitivity, and a *93.67%* specificity.

The results show that the most relevant combination of dynamic criteria is the one based on the features “Relative Enhancement” and “Washout ratio”. This combination reaches *92.5%* mean accuracy, *89.9%* mean sensitivity, and a *95.6%* mean specificity. Furthermore, the multilayer perceptron using this input configuration correctly classifies the two false negative cases present in the dataset.

IV. Conclusion

The main contributions of the experimentations presented in this paper are a new dataset[1] and a thorough experimentation of several dynamic criteria that have been proposed so far for discriminating between benign and malignant patterns in the contrast agent diffusion. We built a very challenging dataset including 1200 real examples of enhancement values which are not easily classified in the two classes. We used the dataset for experimenting several automatic classifiers and their combinations. Model selection is performed via a double k-fold cross validation procedure which allows an accurate and unbiased error prediction. The performance of our models is very good and we were also able to correctly classify two false negative cases. In the near future we plan to include the study of morphological (not dynamic) features in order to improve the performance obtained so far.

References

[1] G. Brix, F. Kiessling, R. Lucht, S. Darai, K. Wasser, S. Delorme, and J. Griebel. Microcirculation and microvasculature in breast tumors: Pharmacokinetic analysis of dynamic MR image series. Magnetic Resonance in Medicine, 52(2):420–429, 2004.

[2] X. Fan, M. Medved, G. S. Karczmar, C. Yang, S. Foxley, S. Arkani, W. Recant, M. A. Zamora, H. Abe, and G. M. Newstead. Diagnosis of suspicious breast lesions using an empirical mathematical model for dynamic contrast-enhanced MRI. Magnetic Resonance Imaging, 25(5):593 – 603, 2007.

[3] P. Filzmoser, B. Liebmann, and K. Varmuza. Repeated double cross validation. Journal of Chemometrics, 23(4):160–171, 2009.

[4] C. I. Mosier. I. problems and designs of cross-validation 1. Educational and Psychological Measurement, 11(1):5–11, 1951.

[5] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. Nature, 323:533–536, 1986.

[6] B. K. Szabó, P. Aspelin, M. Kristoffersen Wiberg, and B. Boné. Dynamic MR imaging of the breast. Acta Radiologica, 44(4):379–386, 2003.

[1] Soon available at the following URL: http://nmis.isti.cnr.it/cardillo/mri/breast-mri-dataset.html