Phys. Med. Biol. **46** (2001) 1651–1663 www.iop.org/Journals/pb PII: S0031-9155(01)21237-8

An SVM classifier to separate false signals from microcalcifications in digital mammograms

Armando Bazzani[1], **Alessandro Bevilacqua**[2], **Dante Bollini**[1], **Rosa Brancaccio**[1], **Renato Campanini**[1], **Nico Lanconelli**[1,3], **Alessandro Riccardi**[1] **and Davide Romani**[1]

[1] Department of Physics, University of Bologna, and INFN, Bologna, Italy
[2] Department of Electronics, Computer Science and Systems, University of Bologna, and INFN, Bologna, Italy

E-mail: nico.lanconelli@bo.infn.it

Received 22 January 2001, in final form 22 March 2001

Abstract
In this paper we investigate the feasibility of using an SVM (support vector machine) classifier in our automatic system for the detection of clustered microcalcifications in digital mammograms. SVM is a technique for pattern recognition which relies on the statistical learning theory. It minimizes a function of two terms: the number of misclassified vectors of the training set and a term regarding the generalization classifier capability. We compare the SVM classifier with an MLP (multi-layer perceptron) in the false-positive reduction phase of our detection scheme: a detected signal is considered either microcalcification or false signal, according to the value of a set of its features. The SVM classifier gets slightly better results than the MLP one (Az value of 0.963 against 0.958) in the presence of a high number of training data; the improvement becomes much more evident (Az value of 0.952 against 0.918) in training sets of reduced size. Finally, the setting of the SVM classifier is much easier than the MLP one.

1. Introduction

Breast cancer is the most common form of cancer among women. The presence of microcalcifications in breast tissues is one of the main features considered by radiologists for its diagnosis. CAD (computer aided diagnosis) systems have been examined in order to assist doctors: the computer output is presented to radiologists as a *second opinion* and can improve the accuracy of the detection. Several techniques developed for the automated detection of microcalcifications can mainly be grouped into three different categories: multiresolution analyses (Yoshida *et al* 1994, Lado *et al* 1999), difference-image techniques (Chan *et al* 1987) and statistical methods (Karssemeijer 1993, Gurcan *et al* 1998, Poissonier *et al* 1998).

[3] Address for correspondence: Department of Physics, Viale Berti-Pichat 6/2, 40127 Bologna, Italy.

By comparing the different methods it turns out that some microcalcifications are detected by one method but missed by others: this is due to the existence of many types of microcalcification. It is often hard for one single detection scheme to discover different types of signal with various characteristics.

In this paper we propose an approach based on the combination of different detection methods in order to get optimal performance. Yoshida *et al* pointed out that the simultaneous use of two or more techniques might improve the results of an optimized single method (Yoshida *et al* 1996). In our method we combine a multiresolution analysis based on wavelet transform with a filtering method (Belikova and Yaroslavsky 1980) and a Gaussianity statistical test and then perform a logical OR operation on the detected signals before clustering (Bazzani *et al* 2000).

A very critical phase of every CAD system is the FPR (false-positive reduction) step: here a detected signal is considered either microcalcification or false signal, according to the value of a set of its features. It is therefore necessary to set up a classifier which, hopefully, maintains quite all the true detected signals and rejects, at the same time, almost all the false positive signals. Other researchers (Woods *et al* 1993, Zhang *et al* 1996, Edwards *et al* 2000) have shown that the use of classifiers based on artificial neural networks can improve the performance of a detection scheme. In this paper we present a classifier based on the SVM (support vector machine).

SVMs have been introduced as a technique which relies on statistical learning theory (Vapnik 1995, 1998). Whereas other techniques, e.g. MLPs (multi-layer perceptrons), are based on the minimization of the empirical risk, that is the minimization of the number of misclassified vectors of the training set, SVMs minimize a functional which is the sum of two terms. The first term is the empirical risk; the second term (confidence term) controls the ability of the machine to learn any training set without error. SVMs are attracting increasing attention because they rely on a solid statistical foundation and appear to perform quite effectively in many different applications (Lecun *et al* 1995, Osuna *et al* 1997, Pontil and Verri 1998). After training, the separating surface is expressed as a certain linear combination of a given kernel function centred at some of the data vectors (named *support vectors*). All the remaining vectors of the training set are effectively discarded and the classification of new vectors is obtained solely in terms of the support vectors.

The aim of our work is to investigate the feasibility of using an SVM classifier in the FPR phase of our CAD detection method and to compare the SVM classifier to the MLP one. Common sets of training data and test data are used to evaluate and compare the classifiers. The performance of the detection scheme has been tested on 40 digitized mammograms from the Nijmegen hospital: this database is considered as a benchmark for CAD systems. The images have been digitized to a pixel size of 0.1×0.1 mm^2 and quantized to 12-bit grey scales.

2. Methods

2.1. Overview of the detection scheme

Microcalcifications are very small spots that are relatively bright compared with the surrounding normal tissue. Typically they are between 0.1 mm and 1 mm in size and are of particular clinical significance when found in clusters of five or more in a 1 cm^2 area. Most of the clusters consist of at least one evident microcalcification and other more hidden signals. Our approach includes two different methods: the first one (coarse) is able to detect the most obvious signals and uses filtering techniques and Gaussianity tests, while the second one (fine), based on multiresolution analyses, discovers more subtle microcalcifications.

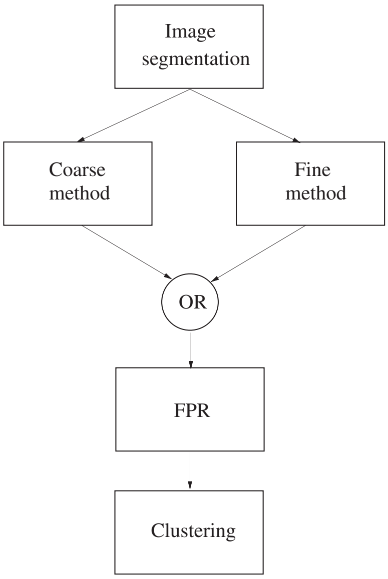

Figure 1. Detection scheme.

First the digitized image is segmented to isolate breast tissues from image background. In this way we reduce both the processing time and memory requirements, since we analyse only areas which contain useful information for the detection. The segmented image is then passed to the two signal-extraction methods described in the following subsections. Signals from these methods are combined through a logical OR operation and then passed to the FPR step. FPR is a two class pattern recognition problem: here the classifier (SVM or MLP) separates true microcalcifications from false signals. The FPR phase is based on a local edge-gradient analysis: we consider five features (area, average pixel value, edge gradient, degree of linearity and average local gradient), which are the inputs of the classifiers. These features are common and often used in microcalcifications detection methods, since they are very useful in discriminating microcalcifications from false-positive signals (Ema *et al* 1995). Finally, signals survived to the FPR phase are clusterized to give the final result. The detection scheme is shown in figure 1.

2.2. Coarse method

In this part of the algorithm we remove structured image background by means of a filtering technique. The scheme of the coarse method is shown in figure 2.

First of all we perform an iso-precision noise equalization as described in Karssemeijer (1993). The equalized image is passed through a linear filter:

$$x'_{i,j} = \frac{1}{(2N_1+1)^2}\sum_{n,m=-N_1}^{N_1} g1_{n,m}x_{i+n,j+m} - \frac{1}{(2N_2+1)^2}\sum_{n,m=-N_2}^{N_2} g2_{n,m}x_{i+n,j+m}$$

where $(2N_1 + 1)$ and $(2N_2 + 1)$ are the sides of the masks $g1$ and $g2$, $x_{i,j}$ and $x'_{i,j}$ are the grey values of the pixel (i, j), respectively before and after filtering; $g1$ and $g2$ are defined according to figure 3.

According to experimental evidences we assume that the remaining noise is Gaussian, since we have reduced the structured noise in the filtering step. We then employ a Gaussianity

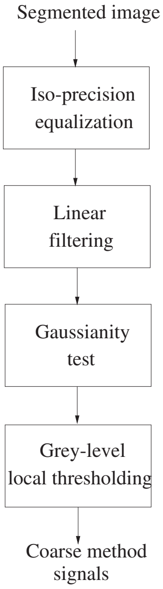

Figure 2. Scheme of the coarse method.

$$g1 = \begin{bmatrix} 0.75 & 0.75 & 0.75 \\ 0.75 & 1 & 0.75 \\ 0.75 & 0.75 & 0.75 \end{bmatrix} \qquad g2 = \begin{bmatrix} 1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1 \\ 1&1&0&0&0&1&1 \\ 1&1&0&0&0&1&1 \\ 1&1&0&0&0&1&1 \\ 1&1&1&1&1&1&1 \\ 1&1&1&1&1&1&1 \end{bmatrix}$$

Figure 3. Filter masks $g1$ and $g2$.

test on the filtered image in order to choose ROI's that include interesting signals. Since this image contains only Gaussian noise and signals with a high contrast we should have a deviation from Gaussianity in regions including microcalcifications. Here we perform the grey-level local thresholding: the central pixel of the considered window of the filtered image is retained only if its grey level is greater than the mean pixel value plus a preselected k multiple of the standard deviation σ; both the mean pixel value and σ are estimated locally inside the window. These signals will join others from the fine method described in the next section.

2.3. Fine method

In this part of the detection scheme we try to discover more subtle microcalcifications, by means of a multiresolution analysis based on the wavelet transform. In figure 4 the scheme of the fine algorithm is depicted.

Microcalcifications are characterized by well defined range size and high local contrast, so we find signals having these features. We split the algorithm into two independent sections.

The first one detects signals having size smaller than 1 mm by means of a multiresolution analysis based on the wavelet transform: we reconstruct the image using the first three scales. To extract interesting signals we perform a local thresholding in 40 × 40 pixel size windows. Assuming for the noise a Gaussian distribution, we fit with a parabola the grey-level histogram

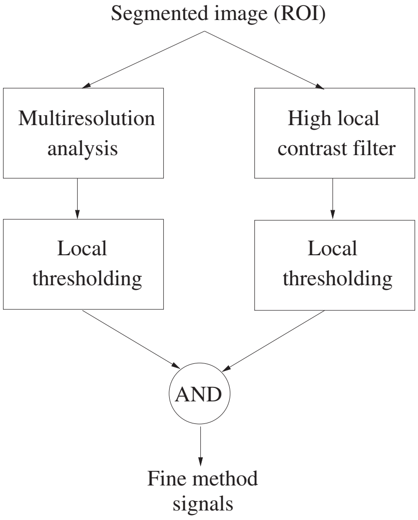

Figure 4. Scheme of the fine method.

of the window: then we retain pixels having a grey-level value greater than the one intersecting the parabola and the x axis.

Signals having a high local contrast are enhanced in the second section, by using a filtering technique. We subtract the image obtained by a 9×9 moving average filtering from the enhanced image from a 3×3 Gaussian filter. We carry out the same local thresholding on the filtered image, followed by a morphological opening operation. After that, a logical AND operation is accomplished on signals extracted by these two sections of the fine method. Finally, as seen, these microcalcifications are joined with others coming from the coarse method through the logical OR operator.

2.4. Overview of support vector machines

SVMs are learning machines used in pattern recognition and regression estimation problems (Cristianini and Shawe-Taylor 2000). They grow up from statistical learning theory (Vapnik 1995, 1998), which gives some useful bounds on the generalization capacity of machines for learning tasks. The SVM algorithm constructs a separating hypersurface in the input space. It acts as follows:

(i) maps the input space into a higher dimensional feature space through some nonlinear mapping chosen *a priori* (kernel);

(ii) constructs the MMH (maximal margin hyperplane) in this feature space; the MMH maximizes the distance of the closest vectors belonging to the different classes to the hyperplane.

Let S be a set of l vectors $\boldsymbol{x}_i \in \boldsymbol{R}^n$, $(i = 1, 2, \ldots, l)$, in a n-dimensional space. Each vector $\boldsymbol{x}_i$ belongs to either of two classes identified by the label $y_i \in \{-1, 1\}$. If the two classes are linearly separable, then there exists a hyperplane, defined by $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, which divides S leaving all the vectors of the same class on the same side. It can be easily shown that the

MMH is given by the solution to the problem

$$\begin{cases} \text{minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{with } y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geqslant 1 \qquad (i = 1, 2, \ldots, l) \end{cases} \tag{1}$$

where $b/\|\boldsymbol{w}\|$ is the distance between origin and hyperplane. This is a quadratic programming problem, solved by the Karush–Kuhn–Tucker theorem. If we denote by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_l)$ the l non-negative Lagrange multipliers associated with the constraints, the solution to the problem is equivalent to determining the solution of the *Wolfe dual* problem:

$$\begin{cases} \text{maximize } \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)y_iy_j \\ \text{with } \sum_i \alpha_i y_i = 0 \qquad\qquad \alpha_i \geqslant 0. \end{cases} \tag{2}$$

The solution for $\boldsymbol{w}$ reads

$$\boldsymbol{w} = \sum_i \alpha_i y_i \boldsymbol{x}_i. \tag{3}$$

The only α_i that can be nonzero in equation (3) are those for which the constraints of the first problem are satisfied with the equality sign. Since most of the α_i are usually null, the vector $\boldsymbol{w}$ is a linear combination of a often relatively small percentage of the vectors $\boldsymbol{x}_i$. These vectors are termed *support vectors* and they are the only vectors of S needed to determine the MMH. The problem of classifying a new data vector $\boldsymbol{x}$ is now simply solved by looking at the sign of $\boldsymbol{w}\cdot\boldsymbol{x} + b$ with b obtained from the Karush–Kuhn–Tucker conditions (Vapnik 1995).

In the case where the set S cannot be separated by any hypersurface, due to the partial overlapping of the two classes, the previous analysis can be generalized by introducing l non-negative *slack* variables $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_l)$ such that

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geqslant 1 - \xi_i \qquad (i = 1, 2, \ldots, l). \tag{4}$$

The solution to

$$\begin{cases} \text{minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \xi_i \qquad (i = 1, 2, \ldots, l) \\ \text{with } y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geqslant 1 - \xi_i \qquad (i = 1, 2, \ldots, l) \end{cases} \tag{5}$$

is called the SMSH (soft margin separating hyperplane). Once again, the vectors satisfying the constraints above with the equality sign are termed *support vectors* and are the only vectors needed to determine the decision surface. Similarly to the linearly separable case, the dual formulation requires the solution of a quadratic programming problem with linear constraints:

$$\begin{cases} \text{maximize } \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)y_iy_j \\ \text{with } \sum_i \alpha_i y_i = 0 \qquad\qquad 0 \leqslant \alpha_i \leqslant C. \end{cases} \tag{6}$$

In figure 5 is depicted an example of a set of nonseparable vectors belonging to two classes A_1 and A_2 (squares and circles), the SMSH H which separates them and the *support vectors*.

The entire construction can be extended rather naturally to include nonlinear separating hypersurfaces. Each vector $\boldsymbol{x}$ in input space is mapped into a vector $\boldsymbol{z} = \Phi(\boldsymbol{x})$ in a higher dimensional feature space. We can then substitute the dot product $\langle\Phi(\boldsymbol{x}), \Phi(\boldsymbol{y})\rangle$ in feature space with a nonlinear function $K(\boldsymbol{x}, \boldsymbol{y})$, named the *kernel*. Conditions for a function to be a kernel are expressed in a theorem by Mercer (Vapnik 1995). Admissible kernel functions are for example the polynomial kernel of dth degree $K(\boldsymbol{x}, \boldsymbol{y}) = (1 + \boldsymbol{x}\cdot\boldsymbol{y})^d$ or the Gaussian

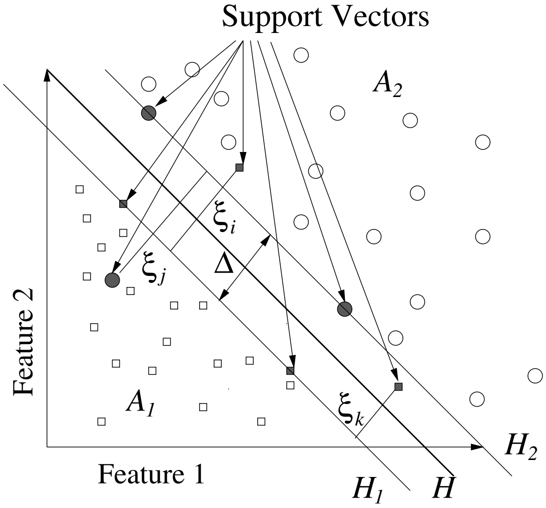

Figure 5. Example of a set of nonseparable vectors belonging to the two classes A_1 (squares) and A_2 (circles). Also depicted are the SMSH H and the two hyperplanes H_1 and H_2, with a distance from H equal to $\frac{1}{2}\Delta$, where $\Delta = 2/\|\boldsymbol{w}\|$ is the margin. Here the *support vectors* (full squares and full circles) are those vectors with distance $\frac{1}{2}\Delta$ from the SMSH and the misclassified vectors.

kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/2\sigma^2)$. Since in the dual formulation example vectors are present only in dot products, performing point (i) becomes quite simple.

We would like to stress here that SVM in the form (5) does suffer from a limitation in two common situations: it is unsuitable both in the case of unbalanced distributions, and when we need to outweigh misclassified examples of one class (e.g. when one type of misclassification is more serious than another). In order to generalize the SVM algorithm to these cases it is necessary to modify (5) in the following way (Morik *et al* 1999):

$$\begin{cases} \text{minimize } \dfrac{1}{2}\|\boldsymbol{w}\|^2 + C^- \sum\limits_i \xi_i^- + C^+ \sum\limits_i \xi_i^+ \\ \text{with } (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geqslant 1 - \xi_i^+,\ (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \leqslant -1 + \xi_i^- \end{cases} \tag{7}$$

where the first sum is for i with labels $y_i = -1$ and the second sum is for i with labels $y_i = +1$ and C^- and C^+ give different costs to false-positive and false-negative errors respectively.

2.5. *Cross-validation of the classifiers*

The combination of the two detection methods described in the previous subsections provides, for a certain configuration of parameters, about 9000 detected signals on the 40 images of the Nijmegen database. Most of them (about 8300) are false-positive signals, whereas only 8% are true microcalcifications. In Nijmegen database we know the ground truth relative to the clusters, but we do not have information about the location of the single microcalcifications inside the cluster. In order to define true and false signals, we have shown the images to three different radiologists, who have marked the true microcalcifications. A detected microcalcification is then defined as true if it is among the signals identified by the radiologists; otherwise it is considered a false-positive. These 9000 signals represent the data on which the classifiers are trained and tested. For each signal a set of five features has been calculated during the detection task, therefore each input for the classifiers is a five-dimensional vector.

The detected signals are divided into three groups: training, validation and test. The first two groups are used to choose the best architecture of the classifier, while through the test group we evaluate its performance on unknown cases. Each group consists of about one third of the total signals and within them the two classes are unbalanced (false signals are about 12 times the true microcalcifications). The problem of having classes with different *a priori* probability is often encountered. For the training of the MLP classifier, we select an equal number of samples from each of the two classes from the training group: we keep all the true microcalcifications and we randomly chose an equal number of false signals. Following Tarassenko (1998) we then perform a post-scaling, in order to reduce the bias towards the more common class. In practice, we scale the output of the MLP after training by a factor equal to the unbalancing rate. Other researchers (Lawrence *et al* 1998) have investigated these issues and discussed different methods for dealing with neural network classifiers in practical situations. We want to stress that the SVM does not require balanced classes, if we setup a classifier following the form (7): in this way it is not necessary to artificially sample the training set. The validation and the test groups are kept unbalanced. We have randomly divided the 9000 detected signals into three groups for nine different times. In the training we have investigated different configurations of classifiers, both in MLP and in SVM cases. By averaging the results over the nine validation groups, we have thus chosen the best MLP and SVM architectures, which has been tested on the nine test groups, in order to give the average performance. We have compared the results of SVM and MLP with an LDA (linear discriminant analysis) classifier: LDA is very easy to use and it does not require the setting of any parameters.

We have also investigated the behaviour of the classifiers with respect to a variation of the size of the training set. To this end, we split the database into two halves: a training group and a test group, each one consisting of 50% of the detected signals. Randomly repeating this operation nine times, we get nine training groups and nine test groups. We perform the training of the classifier with the best configuration previously obtained and calculate the average performance on the nine test groups. For each training group we then select different reduced subgroups consisting of a number of signals ranging from 13% up to 50% of the total signals; we then train the classifiers using these subgroups and average the results on the test group.

ROC (receiver operating characteristic) analysis, which is a widely used method for evaluating the performance of a binary decision-making process in the medical community, is employed to estimate the accuracy of the presented classifiers. The ROC curve is a plot of the classifier's TPF (true-positive fraction) versus its FPF (false-positive fraction). Here the FPF is the probability of incorrectly classifying a false alarm as a microcalcification, whereas the TPF is the probability of correctly classifying a true microcalcification as a microcalcification. The area under the ROC curve (named Az) is an accepted way of comparing the performance of different classifiers. In this paper the ROC analysis is performed by means of the ROCKIT program, developed by Metz *et al* (Metz 1986), which generates an ROC curve for the set of points we are examining. The ROC curve also yields a value of Az, which indicates an unbiased estimation of the performance of the classifier being tested.

3. Results

The first issue faced in this work is the choice of the best configuration of both MLP and SVM classifiers. To this end, we train classifiers with different architectures and estimate their performance on the validation groups obtained as described in the previous subsection. We utilize an implementation of the SVM developed by Joachims (1999), the SVM$^{\text{light}}$

Table 1. Average values of Az in the validation group for different SVM configurations. PLM(i) represents a polynomial kernel of ith degree, Gaussian(i) a Gaussian kernel with $\gamma = i$.

SVM configuration	Az
PLM(2)	0.962 ± 0.001
PLM(3)	0.963 ± 0.001
PLM(4)	0.961 ± 0.002
PLM(5)	0.959 ± 0.002
PLM(6)	0.960 ± 0.002
PLM(7)	0.958 ± 0.002
PLM(8)	0.956 ± 0.002
Gaussian(0.01)	0.934 ± 0.002
Gaussian(0.1)	0.948 ± 0.002
Gaussian(0.5)	0.960 ± 0.002
Gaussian(1)	0.962 ± 0.001
Gaussian(2)	0.962 ± 0.002
Gaussian(5)	0.960 ± 0.002

Table 2. Average values of Az on the test group for the best SVM and MLP configurations and LDA classifier. PLM(i) represents a polynomial kernel of ith degree, Gaussian(i) a Gaussian kernel with $\gamma = i$. The best MLP architecture is a two hidden layer network with 5 × 3 × 2 × 1 neurons.

Classifier configuration	Az
SVM—PLM(3)	0.963 ± 0.001
SVM—Gaussian(1)	0.962 ± 0.001
MLP—(5 × 3 × 2 × 1)	0.958 ± 0.002
LDA	0.930 ± 0.002

program, available at http://ais.gmd.de/~thorsten/svm_light. We have examined two different kernel functions: polynomial with degree ranging from 2 to 8 and Gaussian with values of $\gamma = 1/(2\sigma^2)$ ranging from 0.01 to 5. With fixed $C^- = 1000$, we vary the C^+/C^- ratio from 1 to 12 (the unbalancing rate), in order to obtain the different points of the ROC curve. As the ratio C^+/C^- increases, the loss of the true microcalcifications is weighted more and more; in this way, the sensitivity of the detection method is increased, reducing its specificity. The average values of Az in the validation group are shown in table 1.

It turns out that the performances of all the polynomials and of Gaussian kernels with $\gamma = 0.5, 1, 2, 5$ are very similar. We then evaluate the average results of the best two kernels on the nine test groups, getting the values shown in table 2. We therefore select the polynomial kernel of third degree as the most suitable architecture for our problem. It is important to underline that the choice of the kernel and of its parameter (e.g. degree for the PLM and γ for the Gaussian) is not a delicate issue: different kernels with a wide range of parameters give similar results, as we can see in table 1. Thus, we can state that in our case the setting of the SVM classifier is easy, since its performance does not depends strongly on the choice of the kernel type or on its parameter.

In order to establish the best MLP architecture we have inspected networks with different topologies, using the *Rprop* learning algorithm. For this purpose we utilize a freely available program, the SNNS (Stuttgart neural network simulator) package. We train each network with an equal number of samples of the two classes, obtained from the training groups. Actually, for each training set we use all the true microcalcifications and an equal sample of false signals. It is worth mentioning that, for each MLP network, we perform the training step ten times, with different random inizializations, in order to avoid local-minimum traps. It turns out that the

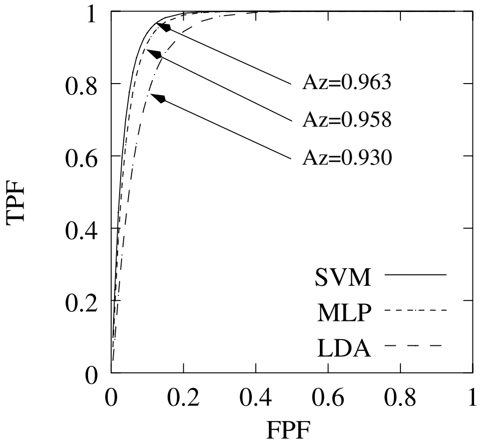

Figure 6. ROC curves on the test group for the best SVM and MLP configurations obtained and for the LDA classifier.

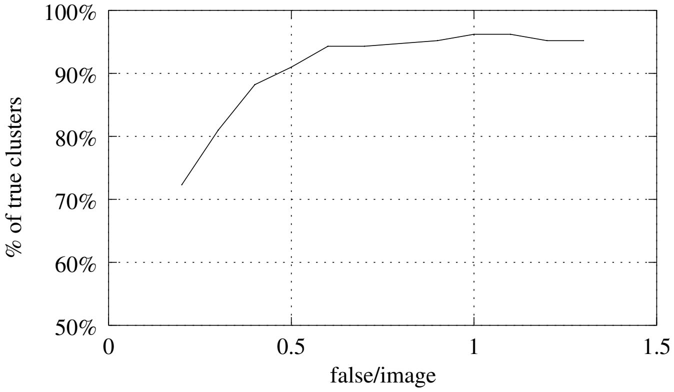

Figure 7. FROC of our detection scheme with the SVM classifier on the 40 images of the Nijmegen database.

best MLP architecture is a two hidden layer network (with 5 × 3 × 2 × 1 neurons) with weight-decay exponent value 6.1, both initial update value and maximum step size equal to 0.33. The different points of the ROC curve are obtained by varying the threshold value of the output neuron. The average values of Az on the test group are shown in table 2. For the LDA classifier we use the LNKnet software, available at http://www.ll.mit.edu/IST/lnknet/index.html.

In figure 6 are depicted three ROC curves relative to the best SVM and MLP configurations and to the LDA classifier. We note that the results of the SVM and MLP classifiers are comparable, whereas LDA gives clearly worse performance. However we want to remark here that the setting of the SVM classifier is much easier that the MLP one: first because there is a reduced number of parameters to be tuned (at most two); second because the SVM acts resolving quadratic problems; consequently it does not suffer from local-minimum traps (in this way it is not necessary to perform training with different random inizializations).

In figure 7 is depicted the FROC (free response operating characteristic) curve, which illustrates the performance of the entire detection scheme with the SVM classifier. We yield a sensitivity of 95% true clusters with 0.6 false-positive clusters per image on the 40 images of the

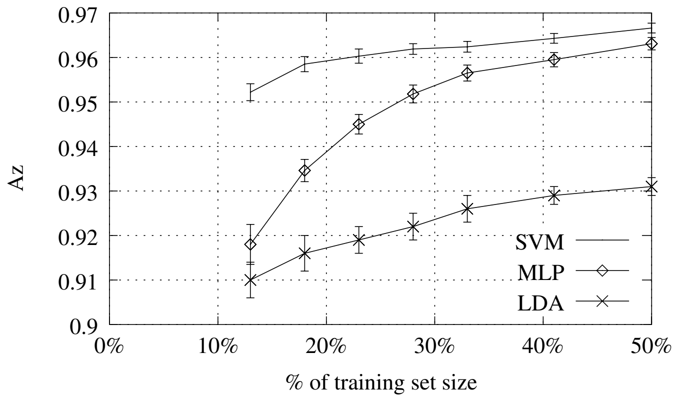

Figure 8. Value of Az as a function of the training set size.

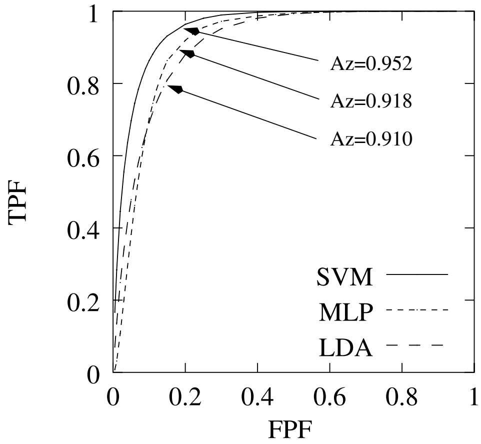

Figure 9. ROC curves on the test group with reduced training test size (about 1000 signals).

Nijmegen database. The curve is relative to the SVM classifier with polynomial kernel of third degree; we calculate the FROC as the average on the whole database of the SVM classifiers trained on the nine training groups already mentioned. Our results are comparable with the best others obtained on the same database (Brown *et al* 1998, Veldkamp and Karssemeijer 1998).

Another issue investigated is the behaviour of the classifiers with training sets of reduced size. To this end we train the best classifiers previously obtained on training groups with different numbers of signals. The size of the training set ranges from 13% (about 1000 signals) up to 50% (4500 signals) of the total detected signals. The test group size is fixed to 50% of the detected signals. The variation of Az as a function of the training set size is depicted in figure 8. We notice that the smaller the training size, the more the SVM outperforms the MLP classifier. This situation is evident in the case of a number of training signals equal nearly to 1000: in figure 9 are depicted the three relative ROC curves on the test groups.

The good performance of the SVM classifier in training sets of reduced size can be extremely useful in several matters, since often it is very difficult to have a large amount of data. We therefore expect to see a more massive use of SVMs, mainly in problems where the scarcity of training data is unavoidable.

4. Conclusion

We have investigated the feasibility of using an SVM classifier in the FPR phase of a CAD method for the detection of microcalcifications in digital mammograms. The results of the entire detection scheme with the SVM classifier are comparable to the best others obtained on the 40 images of the Nijmegen database.

The first advantage of SVM over other traditional classifiers (e.g. MLP) is that its setting is much easier. Besides, SVM does not risk becoming trapped in local minima, since it deals with quadratic problems (hence it always gets to the global minimum). Consequently, for the SVM it is not necessary to repeat the training with different random inizializations. With the SVM classifier we get results comparable with the MLP ones, in any case much better than those obtained with LDA, when the number of training signals is considerably high. On the other hand, the SVM outperforms both the MLP and the LDA classifiers in deficiency of training data.

Therefore we think that SVM classifiers are to be highly recommended for their simple utilization and their good performance, especially in reduced training set size.

Acknowledgments

This work is supported by the Italian National Institute for Nuclear Physics (INFN—CALMA project). Images were provided by courtesy of the National Expert and Training Center for Breast Cancer Screening and of the Department of Radiology at the University of Nijmegen, The Netherlands.

References

Bazzani A, Bevilacqua A, Bollini D, Brancaccio R, Campanini R, Lanconelli N and Romani D 2000 System for automatic detection of clustered microcalcifications in digital mammograms *Int. J. Mod. Phys.* C **11** 901–12

Belikova T P and Yaroslavsky L P 1980 Interactive image preparation in medical diagnosis and natural-resource research *Autometriia* **4** 66–75

Brown S, Li R, Brandt L, Wilson L, Kossof G and Kossof M 1998 Development of a multi-feature CAD system for mammography *Digital Mammography* (Nijmegen: Kluwer) pp 189–96

Chan H P, Doi K, Galhotra S, Vyborny C J, MacMahon H and Jokich P M 1987 Image feature analysis and computer aided diagnosis in digital radiography: automated detection of microcalcifications in mammography *Med. Phys.* **14** 538–48

Cristianini N and Shawe-Taylor J 2000 *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* (Cambridge: Cambridge University Press)

Edwards D C, Kupinski M A, Nagel R, Nishikawa R M and Papaioannou J 2000 Using a Bayesian neural network to optimally eliminate false-positive microcalcification detections in a CAD scheme *Proc. Int. Workshop on Digital Mammography 2000* at press

Ema T, Doi K, Nishikawa R M, Jiang Y and Papaioannou J 1995 Image feature analysis and computer aided diagnosis in digital radiography: reduction of false-positive clustered microcalcifications using local edge-gradient analysis *Med. Phys.* **22** 161–9

Gurcan M N, Yardimci Y and Cetin A E 1998 Microcalcifications detection using adaptive filtering and Gaussianity tests *Digital Mammography* (Nijmegen: Kluwer) pp 157–64

Joachims T 1999 Making large-scale SVM learning practical *Advances in Kernel Methods. Support Vector Learning* ed B Scholkopf, C Burges and A Smola (Cambridge, MA: Mit Press) pp 41–56

Karssemeijer N 1993 Adaptive noise equalization and recognition of micro-calcification clusters in mammograms *Int. J Pattern Recog. Artificial Intell.* **7** 1357–76

Lado M J, Tahoces P G, Mendez A J, Souto M and Vidal J J 1999 A wavelet based algorithm for detecting clustered microcalcifications in digital mammograms *Med. Phys.* **26** 1294–305

Lawrence S, Burns I, Back A, Tsoi A C and Giles C L 1998 Neural network classification and prior class probabilities *Tricks of the Trade, Lecture Notes in Computer Science—State of the Art Surveys* ed G Orr, K R Muller and R Caruana (Berlin: Springer) pp 299–314

Lecun Y *et al* 1995 Comparison of learning algorithms for handwritten digit recognition *Proc. ICANN '95* pp 53–60

Metz C E 1986 ROC methodology in radiologic imaging *Invest. Radiol.* **21** 720–33

Morik K, Brokhausen P and Joachims T 1999 Combining statistical learning with a knowledge-based approach—a case study in intensive care monitoring *Proc. 16th Int. Conf. on Machine Learning*

Osuna E, Freund R and Girosi F 1997 Training support vector machines: an application to face detection *Proc. Computer Vision and Pattern Recognition* pp 130–6

Poissonier M, Highham R, Brady M, Shepstone B and English R 1998 Integration of low-level processing to facilite microcalcification detection *Digital Mammography* (Nijmegen: Kluwer) pp 111–18

Pontil M and Verri A 1998 Object recognition with support vector machines *IEEE Trans. Pattern Anal. Machine Intell.* **20** 637–46

Tarassenko L 1998 *A Guide to Neural Computing Applications* (London: Arnold)

Vapnik V 1995 *The Nature of Statistical Learning Theory* (Berlin: Springer)

——1998 *Statistical Learning Theory* (New York: Wiley)

Veldkamp W and Karssemeijer N 1998 Improved correction for signal dependent noise applied to automatic detection of microcalcifications *Digital Mammography* (Nijmegen: Kluwer) pp 169–76

Woods K S, Doss C C, Bowyer K W, Solka J L, Priebe C E and Kegelmeyer W P Jr 1993 Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography *Int. J. Pattern Recog. Artificial Intell.* **7** 1417–36

Yoshida H, Doi K and Nishikawa R M 1994 Automatic detection of clustered microcalcifications in digital mammograms using wavelet transform techniques *Proc. SPIE* **2167** 868–86

Yoshida H, Doi K, Nishikawa R M, Giger M L and Schmidt R A 1996 An improved CAD scheme using wavelet transform for detection of clustered microcalcifications in digital mammograms *Academ. Radiol.* **3** 621–7

Zhang W, Doi K, Giger M L, Nishikawa R M and Schmidt R A 1996 An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms *Med. Phys.* **23** 595–601