LUNG NODULES DETECTION AND CLASSIFICATION

Paola Campadelli, Elena Casiraghi, Giorgio Valentini

Università degli Studi di Milano
Computer Science Department
via Comelico 39/41, 20135, Milano

ABSTRACT

Image processing techniques and Computer Aided Diagnosis (CAD) systems have proved to be effective for the improvement of radiologists' diagnosis. In this paper an automatic system detecting lung nodules from Postero Anterior Chest Radiographs is presented. The system extracts a set of candidate regions by applying to the radiograph three different and consecutive multi-scale schemes. The comparison of the results obtained with those presented in the literature show the efficacy of our multi-scale framework. Learning systems using as input different sets of features have been experimented for candidates classification, showing that Support Vector Machines (SVMs) can be successfully applied for this task.

1. INTRODUCTION AND MATERIALS

The chest radiography is by far the most common type of procedure for the initial detection and diagnosis of lung cancer, preferred to more sensitive and precise techniques such as MRI or CT; this is due to its non-invasivity characteristics, radiation dose and economic considerations. Several studies (e.g., [1]) list and explain all the factors that may affect the technical production of the radiographic image and its correct diagnostic interpretation. When radiologists rate the severity of abnormal findings, large inter-observer and intra-observer differences occur. Impressive results describing both diagnosticians' error rate and the patients' mortality have been reported by several studies (e.g., [2]), that also demonstrate the potentiality of early diagnosis improvement, suggesting the use of computer programs for radiographs analysis. These are the main reasons why in the last two decades a great deal of research work has been devoted to the study of systems aimed to lung nodules detection, and a wide variety of them have been already proposed and reviewed in [3]. The systems proposed usually start by segmenting the lung area, then they process it in order to increase the visibility (also called conspicuity) of the nodules. Rule based schemes exploiting the main nodule characteristics are then employed to extract all the regions that may contain nodules; since the cardinality of the candidate nodules set thus created is always high, the next step employs rule based system and learning machines to discard the false positives without losing the true nodules. At the state of the art most of the recent experiments employ several different versions of Neural Networks (NN), with different architectures and using different inputs. Since the difficulty at the basis of all the learning schemes presented is the high number of false positives extracted some methods have been presented [4], which extract less candidates but lose too many true positives, leaving the problem open.

In this paper we describe the algorithm and the results obtained by the method used to extract a first set of candidate nodules and classify them. The classification was performed experimenting both NN, with several architectures, and SVMs, with different kernels and different settings of their parameters. Since true and false positives were greatly unbalanced, we applied a cost-sensitive approach to improve the sensitivity of the classifiers. We present only the result obtained with SVMs since they are the most robust and promising.

The method has been tested on a standard database acquired by the Japanese Society of Radiological Technology. It contains 247 radiographs: 154 containing lung nodules and 93 of patients with no disease. The images have 2048×2048 pixels (digitized at a resolution of $0.165mm$ pixel size), and 4096 grey levels. The diameter of the nodules ranges from 5 to $35mm$. All the nodules in the images have been classified according to the difficulties encountered in their detection by the radiologists. They have been divided in 5 classes ranging from obvious to extremely subtle. The algorithms for the segmentation, the enhancement and the candidate extraction work on images down-sampled to a dimension of 256×256 pixels (referred as the *Original Images* in the following). This size has been chosen experimentally to reduce the computational costs without worsening the performances. The features used as input for the learning systems are calculated on the images reduced to 512×512 pixels.

WORK PARTIALLY FINANCED BY PRIN 2004: "NOVEL CLUSTERING TECHNIQUES IN BIOMEDICAL IMAGE SEGMENTATION"

2. CANDIDATE NODULES EXTRACTION

The algorithms described in this section are applied to a lung area segmented by our algorithm, described in [5]. At the state of the art several lung segmentation algorithms have been proposed, but none of them is optimal for the task of lung nodule detection since they do not include in the area of interest the bottom of the chest and the region behind the heart, where lung nodules may still be present. Moreover they are often based on several assumptions about the position and orientation of the thorax in the image. The algorithm developed as the initial step of our system detects both the *visible lung area* (i.e. the one commonly defined), and the parts of the thorax usually excluded (the *not visible lung area*); furthermore it works under no assumption. A detailed description of the results obtained and the comparison with other methods presented in the literature are reported in [5], proving that this is a very good initialization step for a CAD system aimed at lung nodules detection.

The use of a multi-scale framework to extract the nodules is due to the fact that they are characterized by different sizes, different grey levels and contrast characteristics. We think that a multi-scale approach is the missing part of the methods presented in the literature: they are able to highlights nodules with characteristics that belong to a limited range, which is related to the shape of the operators used to enhance and detect them. The scheme developed produces several smoothed version of the *Original Image* by convolving it with gaussian filters whose standard deviation, s, takes values in the range $2-12$, according to the minimum and maximum possible pixel size of the nodule radius. For each scale, s, we then subtract from the *Original image* its smoothed version, to get a resulting *Difference image* where the details visible at that scale are enhanced. Since the distribution of grey levels in a nodule sub-image can be approximated by a gaussian, the result of subtracting to a nodule sub-image its smoothed version is usually an image with a positive peak in the central part of the nodule, and negative values in the neighborhood. Indeed, the histogram of the *Difference Image* shows that most of the pixels take negative values while, on the set of positive values, a peak can always be identified. We create a binary image by selecting all the pixels with a value bigger than the one corresponding to the peak; these pixels correspond to the details that can be identified at the scale s. Summing up all the binary images obtained at different scales we get the *Sum Image* (left of fig.1), where the nodules appear as regions with circular shape of different sizes, characterized by the highest grey levels at the center and surrounded by a much darker ring; these areas are extracted by applying the procedure described below for all the possible radius values $r = [2, 12]$, and combining the results. With the fixed radius r, it calculates for each pixel $P(x, y)$ a coefficient $P_r(x, y)$

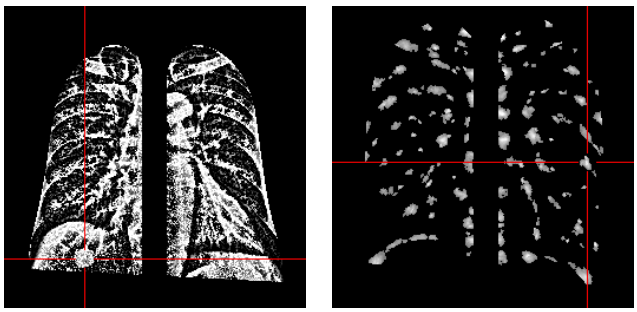

Fig. 1. Sum Image - subtle nodule behind the diaphragm. Regions Image - extremely subtle nodule

which measures the contrast between a circular region with center $P(x, y)$ (and radius r) and its surrounding: $P_r(x, y) = AVG(Circ_r(P(x, y))) - AVG(Rg_r(P(x, y)))$ where $AVG(X)$ is the mean of the grey values of the pixels inside a generic region X; $Circ_r(P(x, y))$ is the region composed by the pixels contained in the circle of radius r and centered in $P(x, y)$; $Rg_r(P(x, y))$ is the region composed by the pixels in the 2-pixel-thick ring around the $Circ_r(P(x, y))$. Note that the thickness of the ring is fixed to 2 for every radius, since what allows to identify a circular region is a darker ring surrounding it, no matter which is the thickness of the ring itself.

To select the pixels which are potential nodule centers, we automatically define a threshold on the set of the coefficients $\{P_r(x, y)\}$, by means of the algorithm described by Kapur in [6], thus obtaining a *binary image*. For each connected region in it, we calculate the circularity and the biggest diagonal, D, of the minimum ellipse containing the region itself. The *circularity* is defined as the fraction of the area of the region contained in the circle with the same area and centered in the center of mass, and the area of the circle itself. We then discard a candidate either if its circularity is lower than 0.5 or D is bigger than $2*r$. The regions left correspond to the candidate nodules with radius r. Repeating the procedure for every possible radius we obtain a set of 11 *binary images* $B(r)$, each containing a set of candidate nodules. All these images are then combined to determine the final set of candidates. First, all the regions appearing in only one of the *binary images* are taken as candidates. For the others the following procedure is employed: when two regions, X and Y, belonging to $B(r_1)$ and $B(r_2)$ (r_1 and r_2 being two consecutive radius values) intersect, their union U is at first considered. If the biggest diagonal D of the minimum ellipse containing it is less than $2 * r_2$, then U is taken as representative; otherwise we calculate the means: $M_X = \frac{1}{|X|} \sum_{p \in X} P_{r_1}$ and $M_Y = \frac{1}{|Y|} \sum_{p \in Y} P_{r_2}$ and take as representative region the one with the higher value. We then build a grey level image, called *Regions im-*

age (right of fig.1), by assigning to each pixel in each candidate region the value $G(x, y) = \max_{r\in[2,12]}(P_r(x, y))$ and then scaling it in the range $[0, 255]$.

With this extraction scheme we get a set of 31100 regions on the 247 images of the database, with an average of about 125 regions per image and only 5 true positives lost out of 154. These results have been compared with those of the extraction schemes tested on the same database and reported in [7] and [8]. The first method has been applied only to the *visible lung area* defined by [9], bringing to a loss of 13 true positives, out of 154, even before the candidate extraction. The result of the extraction scheme is a set of 33000 candidates and a loss of other 8 true positives, for a total of 21 nodules lost at this step. We implemented the second method obtaining really poor results. In terms of number of candidates obtained at this first stage it can be also pointed out that we use a lung area that is about 1.5 times bigger than the one commonly considered, and this has some influence on the cardinality of the obtained set.

To reduce the number of the extracted candidates we calculated several features and perform a statistical analysis to select a set of 16 most representative ones; their efficacy is proved by the fact that their combination by means of simple rules can discard more than 22000 candidates. The drawback of a rule based system are the empirically set thresholds used by the rules, which necessarily bring to a lack of robustness with respect to different databases. For this reason we experimented different learning machines such as NN and SVMs, using as input the same set of 16 features. In the following we will describe just the experiments with the SVMs which gave the most promising results. The set of features computed for each region is composed of:
$\diamond$ the *two* coordinates of its center of mass,
$\diamond$ *three* features describing the position of the region with respect to the *visible* and *not visible lung area* (see [5]),
$\diamond$ *six* features describing its shape,
$\diamond$ the mean grey level of the pixels of the region in the radiograph down-sampled to 512×512 pixels,
$\diamond$ *two* features are the mean and the maximum value of the grey level of the pixels in the *Regions image*,
$\diamond$ *two* different methods [5] have been used to obtain two estimates of the most characteristic radius value associated to the region; they are based on the coefficients $P_r(x, y)$ computed for each pixel and each radius value.

Note that these features are based on the values computed by the extraction scheme; this is due to the observation of the strong dependency between the regions obtained and the algorithm used to extract them. It's a novelty with respect to the methods presented in the literature.

3. CANDIDATES CLASSIFICATION WITH SVMS

The very unbalanced candidate set obtained (30951 False Positives plus 149 True Positives) led us at experiments employing SVMs to discard the biggest number of False Positives. In this context we indeed need to obtain a high sensitivity in order to detect all the positive examples, without a significant loss in specificity, because from a medical point of view it is crucial to detect all the Positive examples, but at the same time we need to significantly reduce the number of False Positives. For these reasons all the results were judged on the basis of their sensitivity and specificity. With SVMs, lowering the decision threshold, we may increment the sensitivity at the possible expense of a decreased specificity. Hence, to better understand the behaviour of the classifiers, we performed a *ROC* analysis, to jointly evaluate in a synthetic way the sensitivity and specificity of the SVMs.

Considering that the data set is very unbalanced (composed of 149 True Positives and 30951 False Positives), for training and testing positive-enriched data sets were built, by considering separately Positive and Negative examples. We randomly split the available Positive data in 89 examples for training and 60 examples for testing according to a train/test ratio equal to $3/2$. From the set of negative data we extracted without replacement a number of negative examples equal to five times the number of positive data, both for the training and the test set, obtaining respectively $89\times5 = 445$ negative examples for the training set and $60\times5 = 300$ negative examples for the test set. We randomly repeated the above process 10 times, obtaining 10 pairs of training and test sets and we normalized the components of the data vectors to 0 mean and unitary standard deviation. We experimented SVMs with linear, polynomial and gaussian kernels, varying the regularization C parameter between 0.001 and 1000, the degree in polynomial kernels between 2 and 6 and the "width" (σ parameter) in gaussian kernels between 0.01 and 10000. Using the ten pairs of training and test sets, we computed the mean and standard deviation of the error, as well as the sensitivity and specificity with respect to the test sets. The experiments generally gave poor results with all the models: even if the average test error generally obtained is quite good (on the average it's equal to 0.11) and the specificity is high (on the average it's equal to 0.96), the corresponding sensitivity is very low (between 0.39 and 0.49 in the models with the lowest test error). Moreover, the best sensitivity value achieved is equal to 0.53, with a value of the specificity slightly lower. We run the same experiments but using different and more complicate sets of features and obtained even worse results. In order to understand the reasons why the SVMs failed to separate positive from negative examples, we analyzed the real-valued discriminant function computed by the SVMs and we ranked the outputs of the SVMs on the test data. In this way low-

ering the threshold of the corresponding decision function we may increment the sensitivity at the expense of a lower specificity. Unfortunately the specificity obtained with a threshold set to obtain a sensitivity equal to 1 is very low: about 0.11 on the average in the best case, and in most cases it also lower. This fact means that the ranking of many positive examples is very low; in other words, it seems that the SVMs "strongly believe" that many positive examples are negative. This may be due to an "intrinsic ambiguity" of the data: examples classified as positive or negative, may not significantly differ with respect to the extracted features. This suggested to run new experiments using a bigger amount of training data; note that increasing the cardinality of the training data necessarily causes a larger unbalance between positive and negative examples. Ten pairs of training test sets were formed in the same way as before, with the same number of True Positive data, but using a positive versus negative ratio equal to $1/30$, hence obtaining respectively $89 \times 30 = 2670$ negative examples for the training set and $60 \times 30 = 1800$ negative examples for the test set. Even though, in this case, the SVMs can learn from more examples, the best sensitivity achieved is equal to 0.32, meaning that the training set is probably too unbalanced.

To overcome this problem other experiments were run using the same unbalanced training and test sets, but introducing a cost-sensitive approach to improve the sensitivity of the SVMs. In classification problems the $0/1$ loss function is usually applied, which weighs equally errors on both positive and negative examples. In medical problems the cost of misclassifying positive (diseased) patients is usually higher than misclassifications of negative (healthy) patients. In the framework of the SVM optimization problem we may introduce regularization parameters C_+ and C_- to be able to adjust the cost of misclassifications of false positives versus false negatives (see [10]). In the experiments presented here we fixed $C_- = C$ and $C_+ = C \times C_f$, where C and C_f are respectively the regularization parameter and the cost-factor; we run experiments where C_f was set equal to 2, 5, 10, 20, 50, 100, so that training errors on positive examples outweigh errors on negative examples. We achieved a significantly higher sensitivity with respect to the previous approaches. With relatively low values of C ($C < 0.01$) and quite large values of the cost factor C_f ($C_f \geq 50$), we obtained sensitivity equal or larger than 0.90 and specificity equal about to 0.70.

Fig. 2 shows the ROC curves of cost-sensitive and standard polynomial SVMs for five different splits of the training and test sets: cost-sensitive SVMs show better sensitivity and specificity compared with those of standard SVMs. Similar results are obtained with linear and gaussian kernels (data not shown).

These results are promising, even if probably not sufficient for clinical pre screening of chest radiographs. Their comparison with the results obtained using more complicate sets of features showed that better performances may be obtained using proper selected set of features; to this end we plan to experiment with cost-sensitive SVMs using feature selection methods to extract subsets of more informative features.

4. REFERENCES

[1] Cj Vyborny, "The aapm/rsna physics tutorial for residents: Image quality and the clinical radiographic examination.," *Rad.*, vol. 17, 1997.

[2] Stitik, *Screening for cancer: Chest Radiology*, Miller AB, edition, 1985.

[3] Van Ginneken, Romeny, and Viergever, "Computer-aided diagnosis in chest radiography: A survey," *IEEE Trans. On Med. Imag.*, vol. 20, 2001.

[4] Yoshida, Xu, and Doi, "Computer-aided diagnosis scheme for detecting pulmonary nodules using wavelet transforms," *Proc. SPIE*, vol. 2434, 1995.

[5] Elena Casiraghi, *A Computer Aided Diagnosis System for Lung Nodules Detection in Postero Anterior Chest Radiographs*, Un. degli Studi, Milano, ITALY, 2004.

[6] Kapur, Sahoo, and Woong, "A new method for gray level picture thresholding using the entropy of the histogram," *Comp. Vis. Graph., Imag. Proc.*, vol. 29, 1985.

[7] Schilham and Van Ginneken, "Multi-scale nodule detection in chest radiographs," *Proc. MICCAI*, 2003.

[8] Bilgin and Hiroyuki, "Computerized detection of pulmonary nodules in chest radiographs based on morphological features and wavelet snake model," *Med. Image Analysis*, vol. 6, 2002.

[9] Van Ginneken and Romeny, "Automatic segmentation of lung fields in chest radiographs," *Med. Phys.*, vol. 27, 2000.

[10] Morik, Brockhausen, and Joachims, "Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring," *Proc. ICML*, 1999.

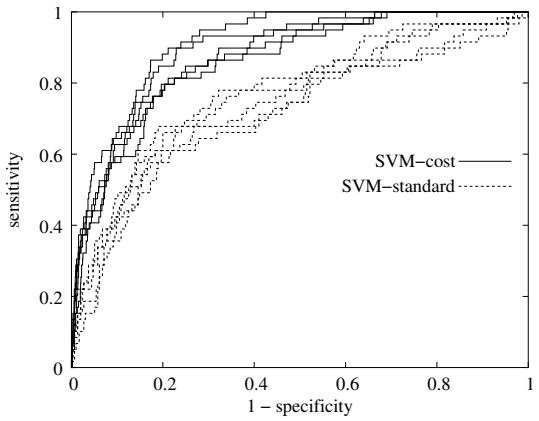

Fig. 2. Comparison of ROC curves in standard and cost-sensitive SVMs.