Diagnosis of Lung Nodule Using the Semivariogram Function

Aristófanés C. Silva¹, Perfilino Eugênio F. Junior²,
Paulo Cezar P. Carvalho², and Marcelo Gattass¹

¹ Pontifical Catholic University of Rio de Janeiro - PUC-Rio
R. Marquês de São Vicente, 225, Gávea
22453-900, Rio de Janeiro, RJ, Brazil

ari@visgraf.impa.br, mgattass@tecgraf.puc-rio.br

² Institute of Pure and Applied Mathematics - IMPA
Estrada D. Castorina, 110, Horto, 22460-320
Rio de Janeiro-RJ, Brazil
perfeuge@visgraf.impa.br, pcezar@impa.br

Abstract. This paper proposes using the semivariogram function, to help characterize lung nodules as malignant or benign in computerized tomography images.

The tests described in this paper were carried out using a sample of 36 nodules, 29 benign and 7 malignant. Fisher's Linear Discriminant Analysis (FLDA), Multilayer Perceptron (MLP) and Support Vector Machine (SVM) were performed to evaluate the ability of these features to predict the classification for each nodule. A leave-one-out procedure was performed to provide a less biased estimate of the classifiers performance. All analyzed classifiers have value area under ROC curve above 0.9, which means that the results have excellent accuracy. The preliminary results of this approach are very promising in characterizing nodules using semivariogram function.

1 Introduction

Lung cancer is known as one of the cancers with shortest survival after diagnosis [1]. Therefore, the sooner it is detected the larger the patient's chance of cure. On the other hand, the more information physicians have available, the more precise the diagnosis will be.

Lung nodules have a structure of very complex tissue. There can be nodules with tissue alterations almost imperceptible to the human eye and other presenting very noticeable alterations. Tissue variation and, sometimes, the not apparent development of the nodule's shape, make diagnosis very difficult. Pattern variations in a nodule's texture (distribution of attenuation coefficients) provide indications about its malignancy or benignity. Nodule calcifications in the shape of a popcorn, laminate concentric, diffuse or central will probably be benign. However, if the nodule does not have calcifications and presents necrosis areas, it is likely to be malignant [1]. The top row in Figure 1 shows the texture for two benign (a and b) and two malignant (c and d) nodules.

Fig. 1. Examples of benign lung nodules and malignant lung nodules.

This work intends to investigate the semivariogram function (a geostatistical function), applied to CT images of three-dimensional nodules and to determine whether they are effective in the diagnosis of lung nodules. The nodule’s malignancy or benignity is determined by applying Fisher’s Linear Discriminant Analysis, Multilayer Perceptron and Support Vector Machine. The validation of the classifiers is done by means of the leave-one-out technique. The analysis and evaluation of tests are done using the area under the ROC curve.

2 Methods

2.1 Image Acquisition

The images were acquired with a Helical GE Pro Speed tomography under the following conditions: tube voltage 120 kVp, tube current 100 mA, image size 512×512 pixels, voxel size 0.67 × 0.67 × 1.0 mm. The images were quantized in 12 bits and stored in the DICOM format [2].

2.2 3D Extraction of Lung Nodules

In most cases, lung nodules are easy to be visually detected by physicians, since their shape and location are different from other lung structures. However, the nodule’s voxel density is similar to that of other structures, such as blood vessels, which makes automatic computer detection difficult. This happens especially when a nodule is adjacent to the pleura. For these reasons, we have used the 3D region-growing algorithm with voxel aggregation [3], which provides physicians greater interactivity and control over the segmentation and determination of required parameters (thresholds, initial and final slice, and seed).

Two other resources help and provide greater control in the segmentation procedure: the barrier and the eraser. The barrier is a cylinder placed around the nodule by the user with the purpose of restricting the region of interest and stopping the segmentation by voxel aggregation from invading other lung structures. The eraser is a resource of the system that allows physicians to erase undesired structures, either before or after segmentation, in order to avoid and correct segmentation errors [4]. The bottom row in Figure 1 shows the 3D reconstruction of the nodules in the top row and exemplifies the nodule segmentation.

2.3 Semivariogram Function

Semivariance is a measure of the degree of spatial dependence between samples. The magnitude of the semivariance between points depends on the distance between the points. A smaller distance yields a smaller semivariance and a larger distance results in a larger semivariance. The plot of the semivariances as a function of distance from a point is referred to as a semivariogram. The semivariogram function summarizes the strength of associations between responses as a function of distance, and possibly direction [5].

A semivariogram has three main features: its sill, range, and nugget (Figure 2). The sill is the ordinate value at which the semivariogram levels off, that is, its asymptotic value; the range is the distance at which this leveling off occurs, that is, the spatial extent of the structure in the data; and the nugget is the semivariance at a distance 0.0, that is, the intercept. A nonzero nugget can imply either intrinsic variability in the data (the component typically ascribed to “sampling error”), or it might indicate that the sampling was conducted at an inappropriate spatial scale, that is, there is considerable variability at scales smaller than the smallest between-point distance.

Fig. 2. Semivariogram and its main features: range, sill, and nugget.

The semivariogram is defined by

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (x_i - y_i)^2 \quad (1)$$

where h is the lag (vector) distance between the head value (target voxel), y_i , and the tail value (source voxel), x_i , and $N(h)$ is the number of pairs at lag h .

When computing directional experimental semivariograms in 3D, two angles are used to define the direction vector: azimuth and dip. To define the rotation of a vector, we assume the unrotated vector starts in the $+y$ direction. The azimuth angle is the first angle of rotation and it represents a clockwise rotation in the horizontal plane starting from the $+y$ axis. The dip angle is the second angle of

rotation and it represents a downward rotation of the vector from the horizontal plane. Other parameters used for semivariogram calculations as lag space, lag tolerance, direction, angular tolerance, maximum bandwidth are exemplified in the Figure 3.

Fig. 3. Parameters used for semivariogram calculations.

2.4 Classification Algorithms

A wide variety of approaches has been taken towards the classification task. Three main historical strands of research can be identified [6]: statistical, neural network and machine learning. This section give an overview of Fisher's Linear Discriminant Analysis, Multilayer Perceptron and Support Vector Machine based on paradigms cited above.

Fisher's Linear Discriminant Analysis - FLDA: Linear discrimination, as the name suggests, looks for linear combinations of the input variables that can provide an adequate separation for the given classes. Rather than look for a particular parametric form of distribution, LDA uses an empirical approach to define linear decision planes in the attribute space i.e. it models a surface. The discriminant functions used by LDA are built up as a linear combination of the variables that seek to somehow maximize the differences between the classes [7]:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta' x \quad (2)$$

The problem then reduces to finding a suitable vector β . There are several popular variations of this idea, one of the most successful being the Fisher Linear Discriminant Rule. Fisher's Rule is considered a "sensible" classification, in the sense that it is intuitively appealing. It makes use of the fact that distributions that have a greater variance between their classes than within each class should be easier to separate. Therefore, it searches for a linear function in the attribute space that maximizes the ratio of the between-group sum-of-squares (B) to the within-group sum-of-squares (W). This can be achieved by maximizing the ratio

$$\frac{\beta' B \beta}{\beta' W \beta} \quad (3)$$

and it turns out that the vector that maximizes this ratio, β , is the eigenvector corresponding to the largest eigenvalue of $W^{-1}B$ i.e. the linear discriminant function y is equivalent to the first canonical variate. Hence the discriminant rule can be written as:

$$x \in i \text{ if } |\beta^T x - \beta^T u_i| < |\beta^T x - \beta^T u_j|, \text{ for all } j \neq i \quad (4)$$

where $W = \sum n_i S_i$ and $B = \sum n_i (x_i - x)(x_i - x)'$, and n_i is class i sample size, S_i is class i covariance matrix, x_i is the class i mean sample value and x is the population mean.

Multilayer Perceptron: The Multilayer Perceptron - MLP, a feed-forward back-propagation network, is the most frequently use neural network technique in pattern recognition [8], [9]. Briefly, MLPs are supervised learning classifiers that consist of an input layer, an output layer, and one or more hidden layers that extract useful information during learning and assign modifiable weighting coefficients to components of the input layers. In the first (forward) pass, weights assigned to the input units and the nodes in the hidden layers and between the nodes in the hidden layer and the output, determine the output. The output is compared with the target output. An error signal is back propagated and the connection weights are adjusted correspondingly. During training, MLPs construct a multidimensional space, defined by the activation of the hidden nodes, so that the two classes (benign and malignant nodules) are as separable as possible. The separating surface adapts to the data.

Support Vector Machine: The Support Vector Machine (SVM) introduced by V. Vapnik in 1995 is a method to estimate the function classifying the data into two classes [10], [11]. The basic idea of SVM is to construct a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples is maximized. The SVM term come from the fact that the points in the training set which are closest to the decision surface are called support vectors. SVM achieves this by the structural risk minimization principle that is based on the fact that the error rate of a learning machine on the test data is bounded by the sum of the training-error rate and a term that depends on the Vapnik-Chervonenkis (VC) dimension.

The process starts with a training set of points $x_i \in \mathcal{R}^n, i = 1, 2, \dots, l$ where each point x_i belongs to one of two classes identified by the label $y_i \in \{-1, 1\}$. The goal of maximum margin classification is to separate the two classes by a hyperplane such that the distance to the support vectors is maximized. The construction can be thought as follow: each point x in the input space is mapped to a point $z = \Phi(x)$ of a higher dimensional space, called the feature space, where the data are linearly separated by a hyperplane. The nature of data determines how the method proceeds. There is data that are linearly separable, nonlinearly

separable and with impossible separation. This last case be still tracted by the SVM. The key property in this construction is that we can write our decision function using a kernel function $K(x, y)$ which is given by the function $\Phi(x)$ that map the input space into the feature space. Such decision surface has the equation:

$$f(x) = \sum_{i=1}^l \alpha_i y_i K(x, x_i) + b \quad (5)$$

where $K(x, x_i) = \Phi(x) \cdot \Phi(x_i)$, and the coefficients α_i and the b are the solutions of a convex quadratic programming problem [10], namely

$$\begin{aligned} \min_{w, b, \xi} \quad & \frac{1}{2} w^T \cdot w + C \sum_{i=1}^l \xi_i \\ \text{subject to} \quad & y_i [w^T \cdot \phi(x_i) + b] \geq 1 - \xi_i \\ & \xi_i \geq 0. \end{aligned} \quad (6)$$

where $C > 0$ is a parameter to be chosen by the user, which corresponds to the strength of the penalty errors and the ξ_i 's are slack variables that penalize training errors.

Classification of a new data point x is performed by computing the sign of the right side of Equation 5. An important family of kernel functions is the Radial Basis Function, more commonly used for pattern recognition problems, which has been used in this paper, and is defined by:

$$K(x, y) = e^{-\gamma \|x - y\|^2} \quad (7)$$

where $\gamma > 0$ is a parameter that also is defined by the user.

2.5 Validation and Evaluation of the Classification Methods

In order to validate the classificatory power of the discriminant function, the leave-one-out technique [12] was employed. Through this technique, the candidate nodules from 35 cases in our database were used to train the classifier; the trained classifier was then applied to the candidate nodules in the remaining case. This technique was repeated until all 36 cases in our database had been the "remaining" case.

In order to evaluate the ability of the classifier to differentiate benign from malignant nodules, the area (AUC) under the ROC (Receiver Operation Characteristic) [13] curve was used. In other words, the ROC curve describes the ability of the classifiers to correctly differentiate the set of lung nodule candidates into two classes, based on the true-positive fraction (sensitivity) and false-positive fraction (1-specificity).

Sensitivity is defined by $TP/(TP + FN)$, specificity is defined by $TN/(TN + FP)$, and accuracy is defined by $(TP + TN)/(TP + TN + FP + FN)$, where TN is true-negative, FN is false-negative, FP is false-positive, and TP is true-positive.

3 Results

The tests described in this paper were carried out using a sample of 36 nodules, 29 benign and 7 malignant. It is important to note that the nodules were diagnosed by physicians and that the diagnosis was confirmed by means of surgery or based on their evolution. Such process takes about two years, which explains the reduced size of our sample.

There were no specific criteria to select the nodules. The sample included nodules with varied sizes and shapes, with homogeneous and heterogeneous characteristics, and in initial and advanced stages of development.

SPSS (*Statistical Package for the Social Sciences*) [14], LIBSVM [15] and NeuralPower [16] were used to training and classification of lung nodules to FLDA, MLP and SVM, respectively. ROCKIT [17] software was used to compute and compare the area under the ROC curve.

Stepwise discriminant analysis [7] was used to select the best variables to differentiate between groups. These measures were used in the FLDA, MLP and SVM classifiers.

In this study, analytical models for the semivariogram were not used; instead, empirical semivariograms were employed. The measures (variables) extracted, considered as texture signatures, were obtained by computing the semivariogram function for a set of directions: dip (Z) 0° , -45° , and -90° ; for each dip the azimuth (X and Y) is 0° , 45° , 90° , and 135° . The adopted lag separation distance (h) was 1, tolerance angle of $\pm 22.5^\circ$, and tolerance lag of ± 0.45 . The maximum number of lags depends on the dimensions of each image (volume). We have selected the first three and the last lags (h) in a specific direction for each function. These lags were selected because we were interested in verifying slight variations in small distance, but without rejecting the information of larger distances. This way, we had 48 measures ($3 \text{ dips} \times 4 \text{ azimuths} \times 4 \text{ lags}$) for semivariogram function. The GSLIB [18] software was used to perform these calculations.

We use the following parameters in the MLP classifier: one hidden layer with four units, hiperbolic tangent as the activation function, the value of 0.15 for the learning ratio, the value of 0.75 for the momentum. These parameters were determined through empirical tests.

In the classification via SVM a proposed procedure by the authors of LIBSVM [15] was used to obtain the best constants C and γ with a process of 36-fold cross-validation. In our case, $C = 2048.0$ and $\gamma = 0.03125$.

Figure 4 shows the application of experimental semivariograms to the volumes represented by Figures 1(a), (b), (c) and (d). We verify that benign nodules have an higher sill than malignant nodules, and that the initial slope is much more accentuated. The graph analysis shows the presence of greater dispersion in benign nodules than in malignant nodules. Table 1 shows the results of semivariogram function and studied classifiers. Based on the area of the ROC curve, we have observed that all classifiers have value AUC above 0.9, which means results with excellent accuracy [19]. There is not statistically significant difference among ROC curves of the classifiers.

Fig. 4. Semivariogram applied to the example in Figure 1.

Table 1. Analysis of FLDA, MLP and SVM classifiers.

Classifiers	Specificity %	Sensitivity %	Accuracy %	AUC
FLDA	86.7	100.0	88.9	0.926 ± 0.071
MLP	93.1	100.0	94.4	0.970 ± 0.046
SVM	100.0	85.7	97.2	0.995 ± 0.019

4 Conclusion

This paper has presented the semivariogram function with the purpose of characterizing lung nodules as malignant or benign. The measures extracted from semivariogram function were analyzed and had excellent discriminatory power, using FLDA, MLP and SVM to classify and the ROC curve to evaluate the obtained results. Based on these results, we have observed that the number of nodules studied in our dataset is too small to allow us to reach definitive conclusions, but preliminary results from this work are very encouraging, demonstrating the potential for multiple variables used in a pattern classification approach to discriminate benign from malignant lung nodules. Nevertheless, there is the need to perform tests with a larger database and more complex cases in order to obtain a more precise behavior pattern.

Despite the good results obtained only by analyzing the texture, further information can be obtained by analyzing the geometry. As a future work, we propose a combination of texture and geometry measures for a more precise and reliable diagnosis.

Acknowledgments

We would like to thank CAPES and FAPERJ for the financial support, Dr. Rodolfo A. Nunes and his team for the clinical support, and the staff from Instituto Fernandes Figueira, particularly Dr. Marcia Cristina Bastos Boechat, for the images provided.

References

1. Tarantino, A.B.: 38. In: *Nódulo Solitário Do Pulmão*. 4 edn. Guanabara Koogan, Rio de Janeiro (1997) 733–753
2. Clunie, D.A.: *DICOM Structured Reporting*. PixelMed Publishing, Pennsylvania (2000)
3. Nikolaidis, N., Pitas, I.: *3-D Image Processing Algorithms*. John Wiley, New York (2001)
4. Silva, A.C., Carvalho, P.C.P.: Sistema de análise de nódulo pulmonar. In: *II Workshop de Informática aplicada a Saúde, Itajai, Universidade de Itajai* (2002) Available at <http://www.cbcomp.univali.br/pdf/2002/wsp035.pdf>.
5. Clark, I.: *Practical Geostatistics*. Applied Science Publishers, London (1979)
6. Michie, D., Spiegelhalter, D.J., Taylor, C.C.: *Machine Learning, Neural and Statistical Classification*. Ellis Horwood Series in Artificial Intelligence, NJ, USA (1994)
7. Lachenbruch, P.A.: *Discriminant Analysis*. Hafner Press, New York (1975)
8. Duda, R.O., Hart, P.E.: *Pattern Classification and Scene Analysis*. Wiley-Interscience Publication, New York (1973)
9. Bishop, C.M.: *Neural Networks for Pattern Recognition*. Oxford University Press, New York (1999)
10. Haykin, S.: *Redes Neurais: Princípios e Prática*. 2 edn. Bookman, Porto Alegre (2001)
11. Burges, C.J.C.: *A Tutorial on Support Vector Machines for Pattern Recognition*. Kluwer Academic Publishers. (1998)
12. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*. 2 edn. Academic Press, London (1990)
13. Erkel, A.R.V., Pattynama, P.M.T.: Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. *European Journal of Radiology* **27** (1998) 88–94
14. Technologies, L.: SPSS 11.0 for windows. Available at <http://www.spss.com> (2003)
15. Chang, C.C., Lin, C.J.: LIBSVM – a library for support vector machines (2003) Available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
16. Software, C.X.: *Neuralpower professional v. 1.0*. Available at <http://www.geocities.com/neuralpower/> (2003)
17. Metz, C.E.: ROCKIT software. Available at <http://www-radiology.uchicago.edu/krl/toppage11.htm> (2003)
18. Deutsch, C.V., Journel, A.G.: *GSLIB. Geostatistical Software Library and User's Guide*. Oxford University Press, New York (1992)
19. Greinera, M., Pfeifferb, D., Smithc, R.: Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine* **45** (2000) 23–41