EVALUATION OF PULMONARY DISEASE USING STATIC LUNG VOLUMES MEASUREMENT IN PRIMARY CILIARY DYSKINESIA

ONLINE DATA REPOSITORY

Soft computing analysis of data

The Principal Component Analysis (PCA) methodology was applied to display data. PCA is a mathematical linear transformation aimed to scale the data in a three-dimensional uncorrelated space, in order to display statistically meaningful differences among the original variables while retaining as much as possible of the variance present in the original dataset. PCA projects the original data along new uncorrelated variables called principal components, i.e. the directions where the variances of data are maximized. These directions are determined by the eigenvectors of the covariance matrix of the original data corresponding to the largest eigenvalues. The prefix eigen- is adopted from the German word "eigen" for "own" in the sense of a characteristic description. The magnitude of the eigenvalues corresponds to the variance of the data along the eigenvector directions. In fact, for a given *p*-dimensional data set X, the *m* principal components Y_1, Y_2,…, Y*m*, where $1<m<p$, are orthonormal axes evaluated by the *m* leading eigenvectors of the covariance matrix, onto which the retained variance is maximum. The information of the observation vectors is contained in the subspace spanned by the first *m* principal components. Therefore, each original data vector can be represented by its principal component vector with dimensionality.

Data were analyzed with soft computing methodologies. Basic elements of soft computing and the application of intelligent control have been recently introduced. The term soft computing denotes methodologies that seek to integrate arithmetical computing, reasoning and decision making into a framework trading off precision and uncertainty. The methodologies used are fuzzy logic, neural networks (NNs) and genetic algorithms and programming. [1] Soft computing-based models are capable of analyzing complex medical data, exploiting meaningful relationships in a data set to help physicians in the diagnosis, treatment and recognition of the clinical outcomes.

KSOMs are artificial neural networks in which the learning process is unsupervised, i.e. the distributed adaptable parameters of the model are autonomously organized on the data. The learning process produces a two-dimensional discrete representation, i.e. a map, of the input space of the input data. Self-organizing maps are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space. This makes self-organizing maps useful for visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling. Such models were first described as artificial neural networks by the Professor Teuvo Kohonen (2).

KSOM is a two-dimensional ANN-based model able to solve classification tasks exploiting structures in the data through an unsupervised learning process.(3) A KSOM maps the original space into a two-dimensional net of neurons in such a way that close neurons respond to similar signals, in order to solve classification tasks and to find structures in data. KSOMs are unsupervised neural networks, i.e. they exploit similarities of samples apart from the class which they belong. In the unsupervised training process, the synaptic weight vectors of the artificial neurons of the KSOM are adapted by means of the training data set examples in such a way that the KSOM supplies as good a representation as possible of the training data set.

The KSOM is composed by an input layer and an output layer . Let X represents the input vector, W the matrix of weights, and Y the output neurons. During the training of the model, at time t, for each output neuron j, the activation is evaluated by the Euclidean distance, i.e. $Y_j = || W_j(t)-X(t)||$. The neuron with the minimum activation is the winning neuron. The weight wij of a generic neuron i at the time T, for the input vector f is modified as follows: $w_{ij}(T)= w_{ij}(T-1)+\alpha(T)[f(T)- w_{ij}(T-1)]$ where $\alpha(T)$ is the gain coefficient selected in the range (0,1). The response of the KSOM is a boolean vector; each element represents the activation function of a neuron. After the training process, a supervised labelling step is performed. Cluster labels are assigned to the individual artificial neurons. After validation of the KSOM by the data set, performance of the classification task is commonly evaluated using the confusion matrix. In order to check the generalization

capability of the neural network, a 10-fold cross-validation process is carried out. In this work, we fixed a 5x5 neurons KSOM with the parameters $\alpha(T) = 0.8$ and a training of 5000 epochs, which allows to obtain the best performance of the model.

The performances of a KSOM predicting model are assessed by using the confusion matrix, which the generic elements i, j indicate how many times in mean percentage ± SD a pattern belonging to the class i was classified as belonging to the class j. In order to check the generalization capability of the KSOM, a k-fold cross-validation is carried out; each fold consists of randomly selected samples, at least one for each category index was included in each fold.

RESULTS

In order to explore the discriminatory power of spirometry and body plethysmography variables in terms of chest HRCT scores and chronic infection with *Pseudomonas aeruginosa* infection, two KSOM models were identified. A ten-fold cross-validation procedure was applied in order to test the performance of each KSOM model.

The identified model based on FEV_1 alone or in combination with FVC, FEV_1/FVC and FEF_{25-75}, is not able to identify any chest HRCT score. The identified model based on body plethysmography variables, alone or in combination, is able to discriminate focal air-trapping severity, atelectasis and the extent of bronchiectasis HRCT scores. The mean ± SD percentages of the confusion matrices obtained by the analysis of the focal air-trapping severity, the atelectasis and the extension of bronchiectasis are reported in table 1. In particular, the KSOM model correctly identifies focal air-trapping with percentages of correct classifications of 94.5%, 77.3 and 81.0% for class of severity 1, 2 and 3 respectively, using the body plethysmography data; whilst using spirometry the discriminatory power is very low (42.0%, 32.5% and 35.0%). The visual representation of the KSOM model based on body plethysmography for the focal air-trapping classification is shown in Figure 4a. Likewise, the model correctly identifies atelectasis and the extent of bronchiectasis with percentages of correct classifications of 66%, 60% and 78.0%, and 79.%, 59.% and of 80.0%, respectively, using the body plethysmography data; by contrast, using spirometry the discriminatory

power is very low (40.0%, 25.0 and 38.6%, and 23.0%, 28.33% and 55.0%). The visual representations of the KSOM models respectively for the atelectasis and the extent of bronchiectasis classifications based on body plethysmography are shown in Figures 4b and 4c respectively. The model correctly identifies the total score with percentages of correct classifications of 80.3%, 63.3% and 90% using the body plethysmography data; while using spirometry does not allow any discrimination (30.0%, 23.3 and 55.0%). The visual representation of the KSOM model based on body plethysmography for the total score classification is shown in Figure 4d. With regard to the other HRCT variables (severity of bronchiectasis, peribronchial thickening and mucous plugging), the model shows the same low discriminatory power as spirometry. These results demonstrate the high discriminatory power of the body plethysmography as compared to spirometry. The identified model based on spirometry and body plethysmography data, alone or in combination, is not able to identify those chronically infected with Pseudomonas.

1. Zadeh LA. The evolution of systems analysis and control: a personal perspective. IEEE Control Syst 1996;16: 95–8.
2. Kohonen T. Honkela, T. Kohonen network. Scholarpedia; 2007;2:1568. http://www.scholarpedia.org/article/Kohonen_network. (accessed 10 October 2010)
3. Honkela, T. Self-Organizing Maps in Natural Language Processing. Thesis for the degree of Doctor of Philosophy Espoo, Finland 1997. http://users.ics.tkk.fi/tho/thesis/ (accessed 12 October 2010).

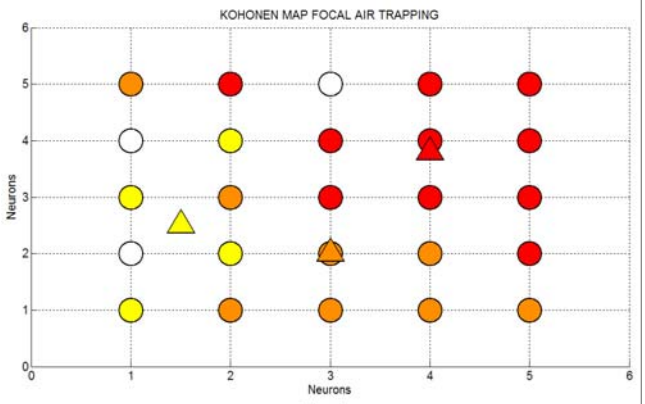

Fig. 4a

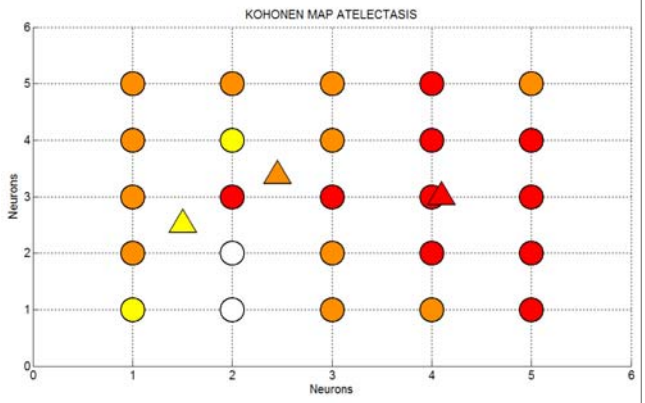

Fig. 4b

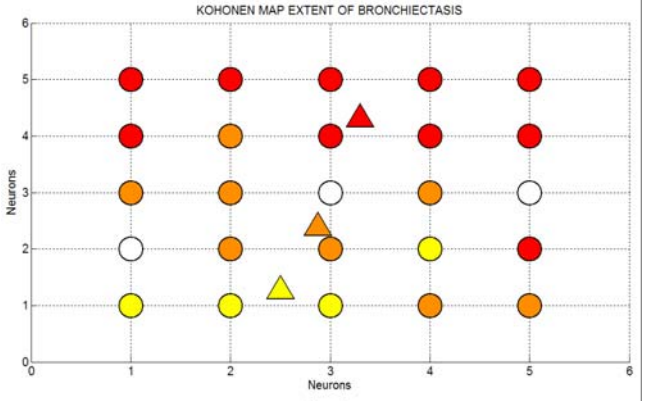

Fig. 4c

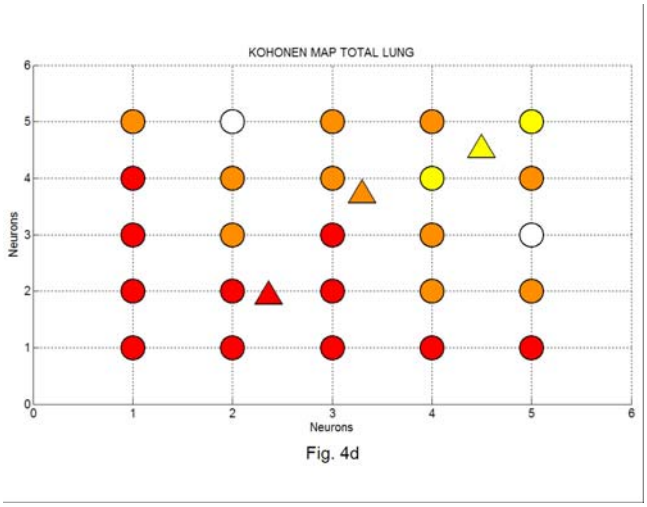

Figure 4 – Visual representations of the KSOMs based on body plethysmography - red: class of severity 1; orange: class of severity 2; yellow: class of severity 3; each triangle indicates the centroid of the corresponding cluster; the inactivated neurons are reported in white; a) focal air-trapping; b) atelectasis; c) the extension of bronchiectasis; d) total score.