

Feature Selection, Ranking Of Each Feature And Classification For The Diagnosis Of Community Acquired Legionella Pneumonia

Enrique Monte¹, Jordi Solé i Casals², Jose Antonio Fiz³, Nieves Sopena⁴.

¹Dpt.TSC.Universitat Politècnica de Catalunya Barcelona.Spain

²Signal Processing Group, Universitat de Vic

³Pneumology dpt, ⁴Infectious Diseases Unit ,Hospital U. Germans Trias i Pujol of Badalona.
E-mail:enric@gps.tsc.upc.es

Abstract. Diagnosis of community acquired legionella pneumonia (CALP) is currently performed by means of laboratory techniques which may delay diagnosis several hours. To determine whether ANN can categorize CALP and non-legionella community-acquired pneumonia (NLCAP) and be standard for use by clinicians, we prospectively studied 203 patients with community-acquired pneumonia (CAP) diagnosed by laboratory tests. Twenty one clinical and analytical variables were recorded to train a neural net with two classes (LCAP or NLCAP class). In this paper we deal with the problem of diagnosis, feature selection, and ranking of the features as a function of their classification importance, and the design of a classifier the criteria of maximizing the ROC (Receiving operating characteristics) area, which gives a good trade-off between true positives and false negatives. In order to guarantee the validity of the statistics; the train-validation-test databases were rotated by the jackknife technique, and a multistarting procedure was done in order to make the system insensitive to local maxima.

1 Introduction

In this paper we deal with the problem of designing a classifier for the diagnosis of Community-acquired pneumonia (CAP), and giving an objective ranking of the importance of the features for classification. The usual tool for determining the importance of a feature is based on the explained variance associated with each feature. This supposes that the multivariate distribution is gaussian, which is not the case when part of the features are qualitative and can take only a limited number of values. In the case of the diagnosis of CAP, most of the features are qualitative, and thus, the use of the variance for explaining the importance of a given feature can be questioned. In this paper instead we use the recognition rate as the criteria for deciding the importance of each feature for classification. Another problem which is dealt in this paper is the selection of a the best subset of features for discriminating the CAP. There are several techniques for finding a good subset d_s of features for classification from a

This research was supported by the CICYT Spanish research project TIC98-0683

given set d . The exhaustive search procedure takes 2^d steps to explore all possible subsets of a given set of size d . If the size of the subset is fixed, there are heuristic search techniques, such as the *branch and bound*, which is efficient for obtaining a subset of a given size, without a combinatorial explosion in the number of steps. This method has the property that is faster than exhaustive search, and yields the subset that maximizes the search criteria, but one must fix the size of the subset from the beginning, and this size might not be the best possible. For this paper we decided to select as search strategy the sequential backward selection, which although does not guarantee that the obtained subset corresponds to the global maximum of the criteria, it gives the eliminated feature at each stage of the search. In the case that there is no strong overlap of the classes in the feature space the method gives the same solution as the exhaustive search in $O(d^2)$ steps. Thus in the process, this search method gives a ranking of the importance of each feature, and the best classifier for the given subset of features. Best in the sense of giving a high recognition rate, (with a low false alarm rate) in the validation and test databases. In order to make the conclusions independent of the given elements in each database (i.e. test, validation and train), we followed the jackknife technique, which consisted in the rotation of the items to be classified through the databases. The problem of local minima of the networks was dealt by the multistarting technique.

Community-acquired pneumonia (CAP) is a common condition throughout the world (1). In Europe, *Streptococcus pneumoniae* is the most frequently identified pathogen, followed by different microorganisms according to the series reviewed. Although Legionella species has been considered as a rare cause of CAP, more recent series have implicated this pathogen in 8% - 15% of the cases (2,3,4). Diagnosis of Legionella pneumonia is commonly performed retrospectively by serology study or tardy by isolation in special culture mediums from clinical samples. Recently, urinary antigen identification (5) has demonstrated high sensitivity (specially if the urine is previously concentrated) (6), and high specificity for detecting *Legionella pneumophila* serogroup 1. This technique is currently being used by several hospitals in many countries to investigate Legionella outbreaks or less frequently as a routine method in the emergency department (7). In spite of its known efficacy, laboratory results are delayed for hours, it is expensive and moreover not all medical centers can perform this technique. However, in patients with CAP, antibiotic therapy cannot be delayed and treatment decisions following different guidelines must be made (8,9). Artificial neural networks (ANN) are tools which have been successfully applied in different medical aspects such as radiologic, clinical and physiologic pattern recognition (10,11,12,13).

2 Methods

Subjects: From May 1994 to June 1999 we prospectively studied 203 patients with laboratory tests diagnosing CAP. Patients over the age of 14 years with acute symptoms consistent with pneumonia and infiltrate on chest radiography at the time of hospital admission or within 24 h were included in the study. Patients with some of the following criteria were excluded: discharge from the hospital less than 10 days

before the onset of symptoms, suspicion of bronchoaspiration, obstructive pneumonia, or pulmonary tuberculosis. Residence in a nursing home, HIV infection, or pharmacologic immunosuppression were not criteria for exclusion. Sixty-two patients had CALP, 141 patients were diagnosed as having community-acquired pneumonia by another bacterial etiology (NLCAP).

Variables studied: For analysis we considered the most predictive variables (table I) according to a previous study (14): (1) demographic variables: age, sex ; (2) risk factors: smoking, chronic alcoholism; underlying disease, including COPD, HIV infection, solid or hematologic neoplasm; pharmacologic immunosuppression (steroids and chemotherapy); (3) variables of clinical presentation: cough, expectoration, thoracic pain, headache, confusion, diarrhea and antibiotic treatment with beta-lactamic drugs prior to pneumonia; (4) analytical data: sodium (Na), creatine kinase (CK) and glutamic oxaloacetic transaminase (GOT).

Definition of the Variables: A patient was considered a smoker if he had smoked >1 pack/day within the last 5 years. Alcoholism was defined as consumption of > 80 gr. of alcohol per day within the same period. Steroid use referred to treatment with >60 mg of prednisone per day over more than 2 weeks in the last month or >5 mg/d for more than the previous 3 weeks. Immunosuppression referred to treatment with cytotoxic drugs not including steroid therapy.

Organization objectives and of the system

We developed the system with two objectives in mind, the first was to provide a set of features selected as a function of their classification potential taking into account the possible interactions between them, and to give a ranking of their importance. This criteria is more realistic than the traditional, based on univariate analysis were using the Student's t test for comparing quantitative variables and the chi square test for qualitative variables, which can be criticized on the basis that it does not take into account possible interactions between features, and also because the hypothesis that assume about the data are not necessary true (i.e., the form of the distribution). Another criticism to this methodology is that the selection of features is done as a function of the level of significance which is indirectly related with the real objective; which is to maximize the compromise between correct classification and false alarms. In order to maximize this compromise explicitly, we took as performance measure of the classifier, the Receiving Operating Characteristic (ROC), which is a measure of this compromise. The ROC chart gives the value of the % of correct classification and % of false alarms for a given threshold for the decision. If the output of the classifier is greater than the threshold, then the class is decided to be present, on the other case, the decision is that the class is not present, thus, the ROC curve gives the dependency of the correct classification and false alarms for the values of the threshold between 0 and 1. In order to have a measure of performance of the classifier independent of the threshold, we took the area under the ROC curve. When this area is 1 (or 100%), the classifier works perfectly independently of the threshold, when the area is 1/2 (50%), the classifier gives the worst performance. For values lower than 1/2 (> 50%), the inversion of the decision rule, improves the performance, which then is transformed into new roc area=100-old roc area.

Thus we took as a classifier a multilayer perceptron, which can approximate an optimal Bayes classifier, and makes few assumptions on the form of the underlying distribution of the data. The structure of the multilayer perceptron (i.e. kind of non-linearities, and number of hidden nodes), was explored in a series of preliminary experiments, in order to establish the general structure. The criteria for selecting the general structure was the generalization capacity of the classifier, using a rotational scheme for the database (i.e. jackknife). The form of the structure was one that used as non-linearity the sigmoid function, and a range of nodes between 5 and 10. This result is consistent with the worst case bound on the generalization capabilities of a network. The rule says that for a fraction of mistakes ϵ the number of weights of a net should be $W \cong N * \epsilon$. In our case, for a generalization rate of 90% ($\epsilon=0.1$), we need about $W \cong 30$ weights. The actual number of weights in the experiments followed a range between was 210 (when all the features were present) and 72 (for the best subset), thus as the classification problem of CAP/NLCAP has a much simpler structure than in the case Baum's study, we can consider that 2 or 3 times the number of parameters of the threshold can guarantee the generalization properties. Also we used for training the multilayer perceptron a conjugate gradient method, which has better convergence properties when the input size of the network is high. This fact was confirmed empirically, and the performance of the network trained by means of a conjugate gradient method was always, (in mean) several points better (in ROC area) than the performance when the training was done by means of the backpropagation with momentum or trained by a second order algorithm such as the Levenberg-Marquart. In order to make the results as much independent of the gradient search as possible, all experiments were done by the multistarting procedure, i.e., 30 and the selection criteria was the ROC area on the validation test.

For the feature selection, we decided to use the *sequential backward elimination*, which is an algorithm, that at a given stage, computes the recognition rate for all the features present at that stage leaving one out. Once this procedure has been done, a new list is created by selecting the subset of features that maximized the ROC area, and the following stage repeats this procedure with this new list. Thus at each stage the selected list consists of the initial list of features minus the feature that when is eliminated the ROC area is maximized (a multistart procedure was done at this point) . As can be seen in figure 1, we obtain the list of the eliminated features that increase the recognition rate until the a maximum is reached, after this maximum is reached, we obtain the list of the features that contribute the maximum to the recognition rate. An alternative to the *sequential backward elimination* is the *sequential forward selection*, which we did not consider because we wanted to take into account the possible interactions between features, in order to make a ranking. The *sequential forward selection*, begins with a list that consists of one feature and at each stage introduces the feature that best increases the performance. Our hypothesis is that in the *backward elimination*, if there is an interaction between features that increases the performance index, they are maintained during the maximum number of stages. In the *forward selection*, depending on the order in which the features are added into the best list, a good combination of features might be lost. This effect does not seem to happen in the backward elimination. The backward elimination is robust to the drop combinations of features that interact depending on the order that the lists are explored at each stage. In table I, we present the best selected lists in three different

experiments (selected as representatives of a set 10 experiments) , and in table II the ROC area obtained in each experiment. The results show that the sensitivity of the system to a different order on the explored features at each stage is low. It is significant that the differences on the variables present at the best subsets of each experiment are related to the last variables to be eliminated in the experiments where these variables is not present in the best list. This can be explained by the fact that at the last two stages before the maximum of the ROC area, the results are within the confidence margin at 95%, (which was of the order or $\pm 2.5\%$). If this is taken into account, the ranking of features obtained in different experiments can be considered consistent, and almost independent of the realization.

Description of the organization of the database for the experiments

The database was divided into three groups (training, validation and testing), and in order to reduce the bias of the estimation of system we used the jackknife technique, with four combinations of the database. This was done by assigning individuals to each base, so that each base was representative of the whole time interval during which the data was collected. A result of this criteria was that each partition of the database was equilibrated in the sense that all had the same proportion of CALP vs. NLCAP. The criteria for defining the training, validation and testing databases was first to organize the database on the basis of the date of the pneumonia, and then to select and alternatively assign individuals to one of four groups. By assigning individuals to each sub-database in this way we guaranteed that no database was biased due to the seasonality of the illness.

3 Results

The study consisted on two phases, a preliminary phase where the potential value features were selected by means of on univariate analysis and a multivariate analysis which selected the best subset of features and yielded the classifier. The main result of the second phase is presented in figure 1 where we show the sequence of eliminated features, as a function of the ROC area. The sequence can be divide in three parts,

- the ascending part, where features that are discarded improve the recognition rate, and thus give us an indication of the least important features by ranking.
- the descending part, where the order in which features are discarded corresponds to their importance to the recognition rate.
- the part around the maximum that corresponds to the maximum ROC area minus the confidence margin (2%), which indicate the uncertainty about the optimal features for classifying the CAP. This uncertainty is reflected in the subsets that are found if the experiment is repeated, changing the order of the features at each stage of the backward selection algorithm, as shown in table I.

As was expected, the system was sensitive to local maxima of the ROC area. Nevertheless this sensitivity was small. As can be seen in table I, the selected subset in three different experiments was similar, and the elements that were eliminated just

before reaching this maxima or just afterward are common to the three experiments. Also as can be seen in table II, the ROC area associated with the different databases is approximately the same for the three experiments. This three experiments were presented as representative of a group of 10 experiments, and we took two samples that were homogeneous on the ROC area and one outlayer in the test results. Although the test results varied from one experiment to another, the order of elimination of variables was quite similar. The high ROC area associated with the validation database, can be explained by the fact, that the criteria for deciding the feature to be eliminated at each stage of the search procedure, was the generalization result on the validation database. Also we used the multistarting technique for selecting a net independent of the local minima of error, and the criteria was also the generalization ability with the validation database. The test database was not used at any point during the whole trained selection procedure.

Discussion.

LCAP pneumonia has a high prevalence among CAP pneumonias being considered as the third highest cause of CAP (18, 4). In most studies early diagnosis of LCAP is difficult because other NLP pneumonias may present with a similar clinical picture and microbiologic CAP tests are not always available or have only retrospective value. In this paper we have presented a tool that might help the diagnosis of LCAP, which has two advantages, it gives a ranking of the importance of the features for classification, and also a classifier that discriminates the LCAP from the NLP.

In a previous study (14) univariate analysis showed that CAP by LP was more frequent in middle-aged, healthy male patients than CAP by another etiology. Moreover, the lack of response to previous β -lactamic drugs, headache, diarrhea, severe hyponatremia, and elevation in serum CK levels on presentation were more frequent in NLCAP by LCAP, while cough, expectoration, and thoracic pain were more frequent in NLCAP by another bacterial etiology. However, in a previous study, multivariate logistic regression analysis only demonstrated significance for underlying disease, diarrhea and elevation in the CK level, although univariate analysis was similar to the present study. We concluded that some clinical and laboratory parameters may be useful to discriminate between LCAP and NLCP, although with a low sensitivity. Based on this study, we have selected the variables which were more relevant in the sample study.

We used a neural network which yields a good sensitivity and false alarm rate, using a maximally discriminating subset of features. This classification system is inexpensive and is easy to perform. The specificity of the antigenuria test by RIA or ELISA is better, but has a lower sensitivity, only detects LP serogroup 1 and needs the support of sophisticated laboratory techniques (6). In contrast, only three parameters of our system (NA, GOT, CK) are routine laboratory techniques. The lower sensitivity and/or time consuming tests of sputum culture and direct immunofluorescence using monoclonal antibodies make these techniques relatively useful in the emergency department. On the other hand serology is a valuable diagnostic test exclusively for epidemiological purposes. As a classification technique the neural networks are a

promising tool, since they may model complex interactions and arbitrary distributions in multivariable data. As a precedent to our study, other studies have been performed in different medical areas, such as the diagnosis of interstitial lung diseases (10), for the analysis of ventilation-perfusion lung scans (12), or in the survival prediction of breast cancer (19). Other studies of breast cancer, concluded that neural nets had better ROC s than radiologic criteria (11). The use of committee nets have also been used in the diagnosis of hepatoma (10). Further uses include prospective validation in the identification of acute myocardial infarction (20). In conclusion, this tool can be used as a complementary aid in medical tasks, due to its simplicity (uses only clinical and common laboratory tests), accuracy, and immediate results.

4 References

1. M. Woodhead. Community-acquired pneumonia guidelines. An international comparison. A view from Europe. *Chest* 1998 .
2. R Blanquer, R Borrás, D Auffal, P Morales, R Menendez, I Subias, L Herrero, J Redon, J Pascual. Aetiology of community acquired pneumonia in Valencia, Spain: a multicentre prospective study. *Thorax* 1991.
3. J Aubertin, F Dabis, J Fleurette, N Bornstein, R Salomon, E Brottier, J Brune, P Vincent, J Miguères, A Jover, C Boutin. Prevalence of legionellosis among adults: A study of community-acquired pneumonia in France. *Infection* 1987.
4. N Sopena, M Sabria, ML Pedro-Botet, JM Montero, L Matas, J Dominguez, JM Modol, P Tudela, V Ausina, M Foz. Prospective study of Community-Acquired Pneumonia of Bacterial Etiology in Adults *Eur J Microbiol Inf Dis* (in press).
5. M. E. Aguero-Rosenfeld, P.H. Edelstein, Retrospective evaluation of the Du Pont radionimmunoassay kit for detection of Legionella pneumophila serogroup 1 antigenuria in humans. *Journal of clinical microbiology*. 1988.
6. JA Dominguez, JM Manterolas, R Blavia, N Sopena, FJ Belda, E Padilla, M Gimenez, M Sabrià, J Morera, V Ausina. Detection of Legionella pneumophila Serogroup 1 Antigen Antigen in Nonconcentrated Urine and Urine Concentrated by Selective Ultrafiltration. *Journal Clinical Microbiology* 1996
7. A.E. Fiore, J.P. Nuorti, O.S. Levine, A Marx, A.C. Weltman, S. Yeager, R.F. Benson, J Prucker, P.H. Edelstein, P Green, Sh R. Zaki, B.S. Fields, J.C. Butler. Epidemic Legionnaires' Disease two decades later: Old Sources, new diagnostic methods. *Clinical infectious diseases*. 1998.
8. M. Woodhead. Management of pneumonia in the outpatient setting. *Seminars in respiratory infections*. 1998.
9. M.S. Niederman. Community-acquired pneumonia. A North American perspective. *Chest* 1998.
10. N Asada, K. Doi, H. MacMahon, S.M. Montner, M. L. Giger, Ch. Abe, Y. Wu. Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases. Pilot study. *Radiology* 1990.
11. J.Baker, P.Kornguth, J.Lo, M.Williford, C.Floyd, Breast Cancer: prediction with artificail neural networks based on BI-RADS standadized lexicon. *Radiology* 1995.
12. J.A. Scott. Neural network analysis of ventilation-perfusion lung scans. *Radiology* 1993.
13. El-Solh AA, MJ Mador,, E. Ten-Brock, DW Shucard, M Abul-Khoudoud, BJ Grant. Validity of neural network in sleep apnea. *Sleep* 1999.
14. N Sopena. M Sabrià-Leal, M.L Pedro-Botet, E padilla, J Dominguez, J Morera, P Tudela. Comparative study of the clinical presentation of legionella pneumonia and other community-acquired pneumonias. *Chest* 1998
15. CH.M. Bishop. Neural Networks for pattern recognition. Clarendon press. Oxford. 1995 .
16. B. Parmanto, P. Munro, Diagnosis of hepatoma by committee. NIPS 94 Post Conference workshop: Neural Network Applications in Medicine, December 1994.

17. . Lippman, L Kukulich, D. Shahian. Predicting the risk of complications in coronary artery bypass operations using neural networks, in G. Tesauro, D. Touretzky and T. Leen (editors) Advances in Neural Information Processing Systems 7, Cambridge, MA: MIT Press.
18. G.D. Fang, M Fine, J Orloff. New and emerging etiology for community acquired pneumonia with applications for therapy: prospective multicenter study of 359 cases. Medicine 1990.
19. H.Burke,D.Rosen, P.Goodman, Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. NIPS*94 Post-Conference Workshop: Neural Network Applications in Medicine, Vail, Colorado, USA (2-3 December 1994).
20. W. Baxt, J Skora. Prospective validation of artificial neural network trained to identify acute myocardial infarction. The Lancet 1996.

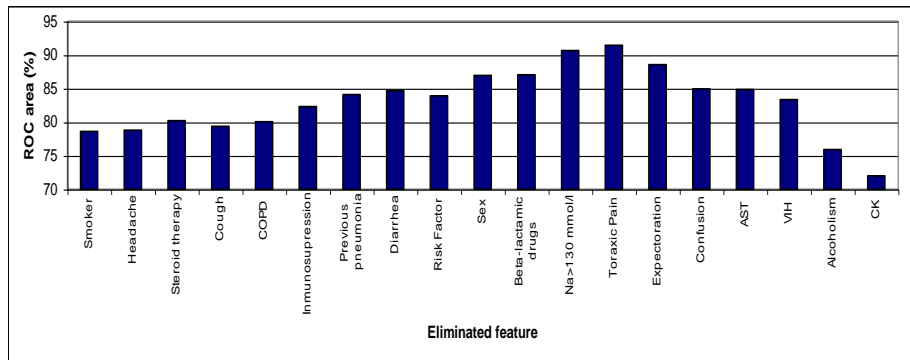


Fig. 1. Sequence of eliminated features as a function of their influence on the ROC area.

Table I Examples on the dependency on the solution with the order of search and the local minima of the neural networks at each step.

List of redundant features (Sorted by stage of elimination)			List of the Best Features		
Experiment 1	Experiment 2	Experiment 3	Experiment 1	Experiment 2	Experiment 3
Sex (m/f) Risk factor Smoker COPD Previous pneumoniae Beta-lactamic drugs Cough Thoracic pain Headache Diarrhea Na>130 mmol/l AST	Sex (m/f) Risk factor Smoker COPD Immunosuppression Steroid therapy Previous pneumoniae Beta-lactamic drugs Cough Thoracic pain Headache Diarrhea Na>130 mmol/l	Sex (m/f) Risk factor Smoker COPD Steroid therapy Beta-lactamic drugs expectoration Thoracic pain Headache Diarrhea AST CK	Age HIV Alcoholism Immunosuppression Steroid therapy expectoration Confusion CK	Age HIV Alcoholism expectoration Confusion AST CK	Age HIV Alcoholism Immunosuppression Previous pneumoniae Cough Confusion Na>130 mmol/l

Table II. ROC area on each database

	Experiment 1	Experiment 2	Experiment 3
Train	83.6 %	80.9	95.1%
Validation	90.5 %	91.4 %	92.3 %
Test	75.7	80.9	70.3%

