Specialized MLP Classifiers to Support the Isolation of Patients Suspected of Pulmonary Tuberculosis

Errison dos Santos Alves, João B. O. Souza Filho
Electrical Engineering Department (DEPEL) / PPEEL
Federal Center of Technological Education Celso Suckow da Fonseca (CEFET-RJ)
Av. Maracanã, 229, Rio de Janeiro, RJ, Brasil
errison_alves@yahoo.com.br, jsouza@cefet-rj.br

Rafael Mello Galliez, Afrânio Kritski
Tuberculosis Academic Program, Medical School,
Federal University of Rio de Janeiro
Rio de Janeiro, RJ, Brazil
galliez77@gmail.com, kritskia@gmail.com

***Abstract*—Tuberculosis is an infectious disease widely present in developing countries, which is largely motivated by the difficulty of a rapid and efficient diagnosis. In order to reduce the number of patients suspected of having TB unnecessarily isolated in hospitals, thus optimize the use of health resources, we propose a systematic procedure for developing a decision support system based on specialized MLP network committee. The system based on 3 MLP models, which response to input data clusters inferred by the k-means technique, exhibits a better classification performance than a single network in terms of the number of false positives, achieving a sensitivity of 83.3% and specificity of 94.3%.**

***Keywords*— Decision Support Systems, Expert Networks, Artificial Neural Networks, Tuberculosis Diagnosis**

I. INTRODUCTION

Tuberculosis (TB) is one of the major infectious diseases that affect the world, particularly in developing countries. The causative agent of TB, *Mycobacterium tuberculosis,* infects the human body via the respiratory tract. Although it may causes disease in various organs of the human body, the lungs usually are the organ most affected, accounting for about 81% of the cases reported in Rio de Janeiro; 25% of them diagnosed in hospitals [1].

The diagnosis of Pulmonary Tuberculosis (PTB) is commonly based on a clinical analysis of patient's signs and symptoms and through diagnostic tests such as baciloscopy and culture. Baciloscopy is a simple and reliable diagnostic test, but has low sensitivity, i.e. exhibits a reduced performance in correctly identifying patients which have the disease. Culture is a more sensitive test, but it takes about 6 to 8 weeks to be concluded and is also restricted to reference or research laboratories. Therefore, since 2006, following the STOP TB Global plan, new diagnostic approaches were proposed by WHO [2].

Thus, fast and efficient TBP diagnosis is a challenge, particularly on the Brazilian National Health System (SUS), demanding an urgent development of new diagnosis approaches. Decision support systems (DSS) based on computational intelligence techniques can provide very useful support tools for tuberculosis diagnosis, especially those involving predictive mathematical models based on artificial neural networks (ANN).

Considering the context of the respiratory isolation of patients suspected of having pulmonary tuberculosis in hospitals, DSS can help medical decision making, reduce the number of unnecessary patient isolations, allow a better management of hospital resources as well as reduce health operational costs [3]. Due to influence life-risk decisions, these systems should exhibit high accuracy and robustness to different epidemiological scenarios.

Usually, classification systems employing multiple models [4] and following the strategy of divide to conquer tend to perform better in complex problems than single models. Thus, this work proposes a systematic procedure to develop a classification system based on multiple models and input data clustering to support pulmonary TB diagnosis. According to this proposal, a reference model synergistically acts together with specialized classifiers derived upon low performance clusters to improve system specificity (number of TB negative patients correctly identified by the system), i.e. to reduce the number of false-positive cases.

The structure of the paper is as it follows: first, we discuss the architecture of specialized models, together with practical aspects related to its production. In the sequence, the dataset is presented and the results are discussed. Finally, we have the conclusions and future work.

II. SPECIALIZED MODELS SYSTEM

The proposed system of specialized models is based on a problem decomposition which follows the strategy of spacial data division [4,5,6], performed through a clustering technique [7]. Some advantages of this approach are: (i) it simplifies the process of obtaining a solution since problem is decomposed in several simpler and smaller ones, (ii) it allows a cluster-to-cluster identification of relevant variables, which may be tuned according to cluster specificities and (iii) it usually results in more robust models, less prone to design issues and data errors.

The proposed architecture is shown in Figure 1. Roughly, for one or more clusters identified by clustering analysis, especially those which exhibit poor classification performance, are produced specialized models. Thus, for data

clustered into *n* groups, each one of the *m* models can be assigned to classify events belonging to one or more clusters. At the end, a supervisory system, based on the same input data clustering, selects the model output to be assigned to any arbitrary input.

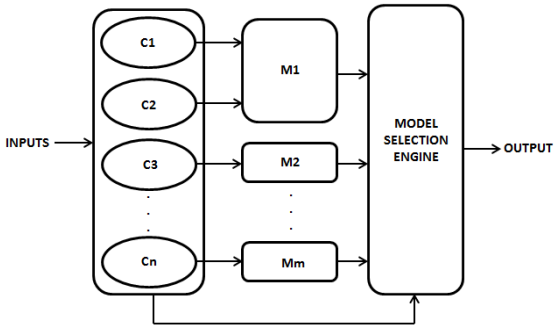

Fig. 1: Architecture of the classification system based on specialized models

A. Development of the specialized models

For the construction of specialized models, an iterative procedure is proposed. Its main objective is to evaluate the adequacy of model specialization, i.e. if the production of models to some input data clusters improves classification performance. Here, the specialization only considered low performance clusters. The procedure is illustrated on Figure 2.

According to it, initially, data is split into clusters and the performance of the reference model is inferred for each group isolatedly. This evaluation can be based on different performance indices. Here, the shape of the ROC curve (qualitative analysis) and values of sensitivity and specificity (quantitative analysis) were considered. Then, specialized models are produced for lower performance clusters and compared to the reference model. If model specialization for a given cluster improves its classification performance, it is inserted into the classification system; otherwise, reference model is assigned to this cluster.

Due to specialization, each new model may consider a different subset of variables to predict the outcome. This particular choice can result in more parsimonious [8] and accurate models. Moreover, the number of events of a problematic cluster must be considered to define if a specialized model will be produced or not. Clearly, this number should permit a proper model deviation and is related to clustering granularity. Fusions of spatially close problematic clusters or involving them and neighbor ones may also be evaluated in these cases.

B. Variables selection

For the selection of input variables used in each neural model, the *wrapper* strategy [7] based on logistic regression model was used. Logistic model is commonly used in medical problems and relates a categorical variable (in this case, binary) with explanatory continuous and/or binary ones (usually considered independent) through a simple non-linear descriptive model, whose parameters are estimated by the EM algorithm [9].

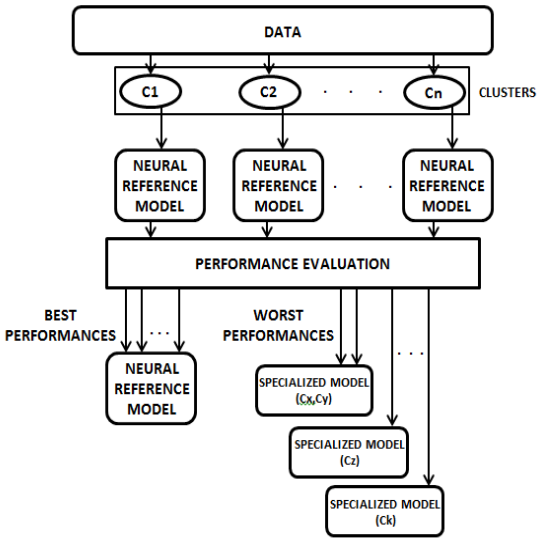

Fig. 2: Illustration of the procedure proposed for the construction of specialized models

Considering the *wrapper* selection, the method of *forward floating search* [7] was adopted. This is an iterative procedure where the model initially has only a constant term, one variable is inserted at each step and this inclusion is defined according to statistical hypothesis tests considering an arbitrary significance level. For all steps, variables outside the current model are evaluated to inclusion and the one which achieves the lowest *p*-value (highest significance for inclusion) is inserted. Additionally, after the inclusion of any variable, the exclusion of one of the current model variables is evaluated, again with basis on hypothesis tests, considering, however, another significance level. In this case, the variable showing highest *p*-value (lowest significance for maintenance) is selected for exclusion. This process continues until no variable can be included or excluded from the model. This selection was performed using *Statistical Package for Social Sciences* (SPSS) software [10] and considered a significance level for variable inclusion of 25% and exclusion of 30%.

C. Data mining

Mining procedures were realized upon dataset, aiming to identify and eliminate events (patients) with inconsistent data which might compromise model learning.

This process excluded patients whose data have gross padding errors or exhibited a percentage of missing values

greater than 25%. Additionally, the exclusion of patients TB+ and TB- with same values for their descriptive variables was realized. These patients were evaluated by TB experts as atypical cases and require a more profound medical investigation to define their correct diagnosis, thus do not fit into the application scope of the proposed system.

Subsequently, aiming to produce better training and test sets, TB+ and TB- patients were divided according to SAFE and DANGER criteria [11,12], which classify events according to their neighborhood. Following these criteria, an event is classified as DANGER if a given number of his nearest neighbors is labeled as belonging to a distinct class and SAFE, otherwise. These strategies are commonly used to identify errors on event labeling or those critical, i.e. which belong to different classes but exhibit high similarity.

D. Dataset

The dataset used in model design consists of clinical information regarding patients admitted to the IDT/HUCFF/UFRJ that were allocated to respiratory isolation rooms due to suspect of pulmonary tuberculosis. These data cover the period of 2001 to 2008.

The database consists of 972 events, with 36 variables related to clinical and symptomatic information. The estimated TB prevalence is of 21.6% (210/972).

III. Results

All neural models were developed using *Multilayer Perceptron* (MLP) [13] networks having 3 layers, one neuron at the output layer and employing a number of hidden layer neurons defined through cross-validation. All neurons considered the hyperbolic tangent as the activation function [14]. The input variables, except age, were coded as: +1, signaling the presence of the sign or symptom; -1, the absence, and 0 when missing. The presence of tuberculosis was coded as +1 and its absence as -1. Patient's ages were normalized to be in -1 to +1 range. Network training used *Resilient Backpropagation* (RPROP) [14] technique.

Hold-out technique was employed to infer model performance with basis on training and test sets formed using SAFE and DANGER subsets. The distribution of SAFE and DANGER events between these sets were defined through some trials guided by cross-validation. Additionally, given the identification of problems concerning dataset class coverage, especially for the more critical patients, the definition of design parameters, including the number of neurons at the hidden layer, also employed the test set.

A. Reference Model

For the construction of the reference model, a selection of variables based on the method described earlier resulted on the following 19 variables: *age, sex, sputum production, hemoptysis, hemoptoic, night sweats, dyspnea, fever, malnutrition, X-ray report (active TB, sequel or another disease), cough, chest pain, HIV, alcoholism, malignancy, smoking and sore throat.*

The amount of SAFE and DANGER events distributed into training and test sets are summarized in Table 1. Note that 17 patients DANGER TB- were excluded after TB expert's criticism (again patients outside the application scope of the score).

TABLE I. Number of Training and Test Sets Events Considering the Reference Model

Events			*Training set*	*Test set*
TB+	*SAFE*	40	-	40 (100%)
	DANGER	121	119 (99%)	2 (1%)
TB-	*SAFE*	640	376 (60%)	264 (40%)
	DANGER	17	-	-
Total		818	495	306

Networks having from 1 to 25 neurons in the hidden layer were built. For each network, 100 trainings with random initial parameters were generated, aiming to avoid the local minima problem [13]. Early stop criterion was also adopted to mitigate the *overtraining* phenomena [13].

The selection of the best model was based on the mean value of the areas between the sensitivity and specificity curves and the thresholds axis for a range of values between -1 and 1. Additionally, the selected model should exhibit a higher value of this performance index for training set than for the test one. The best model had 13 neurons at the hidden layer and its ROC curve is shown in Figure 3. In this plot, the areas under ROC curve (AUC) related to training and test are 0.9807 and 0.9345, respectively.

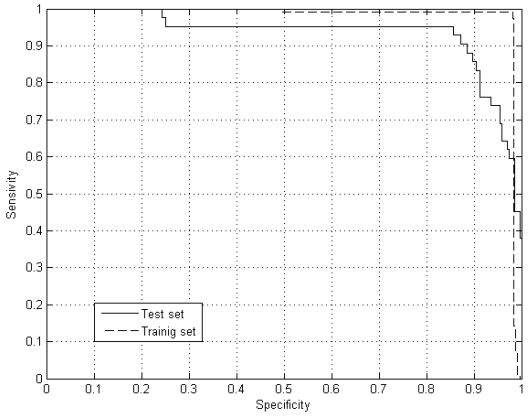

Fig. 3: ROC curves for the reference model (inferred upon training and test sets)

B. Reference model analysis

For reference model performance analysis, clusters were produced by *global k-means* algorithm [15], considering the

Euclidean distance to measure the similarity between events and clusters centers.

Clustering was produced considering all events from dataset, regardless of the outcome (TB+ or TB-) and sets to which they were allocated (training or test). The algorithm *k-means* extracted 15 clusters, since *k* was chosen 15, a value defined empirically.

Empirical cumulative distribution function (CDF) from reference model output is shown in Figure 4 for low performance clusters (8, 9 and 12). This analysis focused only on TB- patients, since the main goal was to improve model specificity, i.e. reduce the number of false-positive cases. It should be mentioned that CDF curves point out specificity values related to different decision threshold choices, thus permit the identification of critical clusters.

Figure 4 shows the existence of several flat regions on the specificity curve, especially for cluster 9 events and decision threshold chosen in the range from -0.8 to 0.8. Thus, improve specificity by tuning this threshold is not possible, which strongly indicates the production of specialized models for these critical clusters.

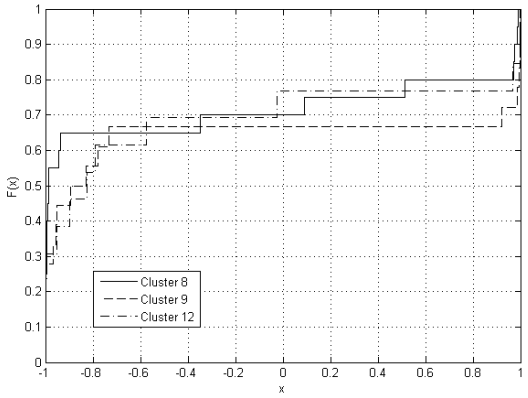

Fig. 4. Cumulated distribution function (CDF) curves of reference model outputs considering low performance clusters.

C. *Identification of clusters candidates to fusion during the production of specialized models*

In order to allow the construction of specialized models involving a larger number of events, thus based on better statistics, the fusion of clusters previously identified as problematic with their neighbors (in spatial sense) was evaluated. The identification of clusters candidates to fusion was guided by a hierarchical clustering derived upon clusters centers. This clustering employed Euclidean distance to measure the similarity between events and mean-linkage criterion to define clusters [16]. Hierarchical cluster dendrogram [16] is shown in Figure 5.

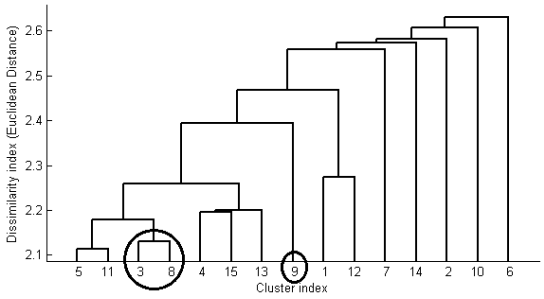

Fig. 5. Hierarchical cluster of *k*-means clusters prototypes

Figure 5 shows higher similarities between the clusters 5-11, 3-8, 1-12 and 4-15-13. Thus, we first evaluated the fusion of clusters 3 and 8. Regarding cluster 9, hierarchical analysis showed no other cluster with a reasonable level of similarity to be fused with it, thus it was kept alone. Other suggested fusions were also evaluated.

D. *Specialized models production and evaluation*

Issues related to the construction and evaluation of specialized models will be described below:

1) *Specialized model for the clusters 3 and 8*

Logistic model identified 12 relevant variables: *age, sex, cough, sputum production, hemoptysis, night sweats, fever, malignancy, HIV, alcoholism and X-ray report (active TB or other disease).* Training and test sets considered the same split performed during reference classifier design, but restricted to groups 3 and 8 events. In Table II, the amount of events destined to each set is shown.

TABLE II. NUMBER OF TRAINING AND TEST SET EVENTS FOR THE SPECIALIZED MODEL DERIVED UPON CLUSTERS 3 AND 8

Clusters 3 and 8			
Events		*Training set*	*Test set*
TB+	33	22	11
TB-	105	61	44
Total	138	83	55

The production of specialized models followed the same framework employed during reference model development. This resulted in a network with two hidden layer neurons.

Figure 6 shows ROC curves for reference and specialized models considering cluster 8 and 9 events (test set). Specialized model achieved higher values of sensitivity when operating with specificity higher than 90%.

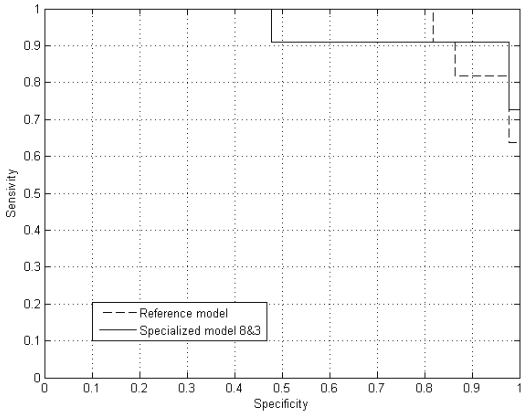

Fig. 6. ROC curves for reference and specialized models considering cluster 8 and 3 events (inferred upon test set).

2) Specialized model for cluster 9

The following 10 variables were selected by logistic method: *age, headache, weight loss, anorexia, liver disease, diabetes, IRC, malnutrition, transplant and RX report (normal).*

Table III shows the amount of events used in training and test sets. The network with better performance had two neurons at the hidden layer. Comparative ROC curves for both reference and specialized model are shown in Figure 7. Again, specialized model achieved better sensitivity values if specificity is set higher than 60%.

TABLE III. NUMBER OF TRAINING AND TEST SET EVENTS FOR THE SPECIALIZED MODEL DERIVED UPON CLUSTERS 9

Cluster 9			
Events		*Training set*	*Test set*
TB+	27	14	13
TB-	37	19	18
Total	64	33	31

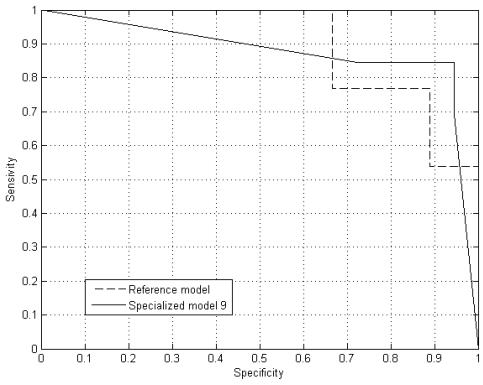

Fig. 7: ROC curves for reference and specialized models considering cluster 9 events (inferred upon test set).

3) Other clusters

Another fusions suggested by hierarchical clustering were evaluated, but none improvement on system performance was observed.

E. Integration of specialized and reference models

The final architecture for the classification system is shown in Figure 8. During the operational phase, for an arbitrary input event, the cluster to which it belongs is identified. If this cluster is other than 3, 8 and 9, the classification is provided by the reference model, otherwise, the output of the corresponding specialized model is considered. ROC curves comparing this classification system with the reference model are shown in Figure 9. It can be observed that the system based on multiple models can achieve better values for sensitivity if specificity is higher than 90%. This fact allows this score to be used as a medical decision support tool with respect to the isolation of patients suspected of having pulmonary TB.

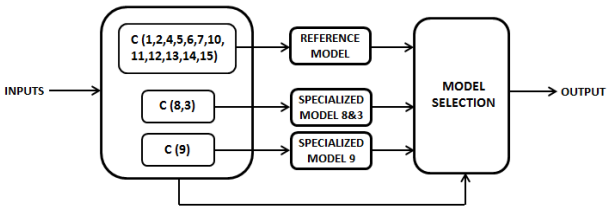

Fig. 8: Final architecture of the classification system based on specialized models

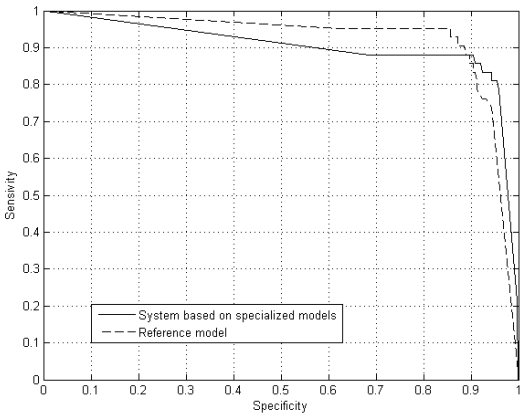

Fig. 9: ROC curves for both reference model and the classification system based on specialized models (inferred upon test set).

IV. Conclusion

This work proposes a classification system based on specialized models to support the isolation of patients suspected of having pulmonary tuberculosis.

Exploring the concept of divide to conquer, we propose a systematic procedure which clusters input data and identify more critical ones in terms of performance. Specialized models are proposed for critical clusters to act in synergistic fashion with the reference model. An interesting aspect is that classifiers can be individually tuned according to specific cluster issues, involving subgroups of the reference model predictive variables, which may result in higher classification performance.

Following the proposed approach, a classification system using 3 neural models (one involving 13 neurons, while others 2 neurons) and 15 clusters achieved a sensitivity of 83.3% to specificity of 94.3%. This represents a gain of almost 10 percentage points on sensitivity when specialized system is compared to reference classifier.

As future work, we intend to explore other computational intelligence algorithms as support vector machines to develop the reference and specialized classifiers as well as to produce classifiers committee. Produced models also will be validated with other datasets.

References

[1] World Health Organization: Global Tuberculosis Control: surveillance, planning, financing. World Health Organization, Geneva, 2010.

[2] WHO 2010 - Tuberculosis (z.d.). Available: http://www.stoptb.org/global/plan/main/default.asp. Accessed on 2013, march.

[3] Souza Filho, J. B. O. ; Vieira, A. P. P. ; Seixas, J. M. ; Aguiar, F. S. ; Mello, F. C. Q. ; Kritski, A. L. An Intelligent System for Managing the Isolation of Patients Suspected of Pulmonary Tuberculosis. In: 13th International Conference on Intelligent Data Engineering and Automated Learning, 2012, Natal. 13th International Conference on Intelligent Data Engineering and Automated Learning, 2012. v. 1. p. 1-8.

[4] Kuncheva, L. I., Combining Pattern Classifiers: Methods and Algorithms, New Jersey, John Wiley & Sons, 2004..

[5] Rokach, L., Pattern Classification Using Ensemble Methods, World Scientific, 2010.

[6] Sharkey, A. J. C., Combining Artificial Neural Nets: Ensemble and Modular Multi-net Systems, Springer-Verlag, 1998.

[7] Theodoridis, S., Koutroumbas, K., “Pattern Recognition”, 4th ed., Elsevier, 2009

[8] Medeiros, M., Terasvirta, T., Rech, G., “Building Neural Networks Models for Time Series: A Statistical Approach”, Journal of Forecasting, v. 25, pp. 49–75, 2006.

[9] Hosmer, D. W., Lemeshow, S., Applied Logistic Regression, 2nd ed., John Wiley & Sons, 2000.

[10] Levesque, R., SPSS Programming and Data Management - A Guide for SPSS and SAS Users, 4th edition, SPSS Inc., 2007.

[11] Bunkhumpornpat, C., Sinapiromsaram, Lursinsap, C., "Safe-Level-SMOTE: Safe-Level Synthetic Minority Over-sampling Technique", Advances in Knowledge Discovery and Data Mining, vol. 5476, 2009, pp. 475-482.

[12] Han, H., Wang, W. Y., Mao, B. H.; "Borderline-SMOTE: A new Over-Sampling Method in Imbalanced Data Sets Learning", Proc. Int'l Conf. Intelligent Computing, 2005, pp. 878-887.

[13] Haykin, S., “Neural Networks: A Comprehensive Foundation”, 2nd ed. New Jersey, Prentice-Hall, 2008.

[14] Riedmiller, M., Braun, H., “RPROP: A Fast Adaptive Learning Algorithm”, International Symposium on Computer and Information Science VII, Antalya, Turkey, 1992, pp. 279-286.

[15] Likas, A., Vlassis, N., Verbeek, J. J., "The Global k-means Clustering Algorithm", Pattern Recognition, vol. 36, 2003, pp. 451-461.

[16] Kaufman, L., Rousseeuw, P. J. (eds.), Finding Groups in Data - An Introduction to Cluster Analysis. John Wiley & Sons, 2005.