8

A Decision Support System Based on Artificial Neural Networks for Pulmonary Tuberculosis Diagnosis

Carmen Maidantchik et al.*
Signal Processing Lab, COPPE/POLI, Federal University of Rio de Janeiro
Brazil

1. Introduction

In 2005 the Faculty of Medicine, the Electronic and Computing Engineering Department of the Federal University of Rio de Janeiro (UFRJ) and the Electrical Engineering Department of the Federal Center of Technological Education (CEFET-RJ) started a collaborative research project to develop a Decision Support System for Smear Negative Pulmonary Tuberculosis (SNPT). The motivation was to develop, through a multi-disciplinary, multi-institutional, innovative and cost-effective approach, new paradigms to prevent the disease progression and support the rapid evaluation of new therapies. The project also aims at increasing the scientific and technological capacity in the country for the progress of new technologies incorporation in the public and private system as well revising public policies to control tuberculosis.
The conception of the initiative was based on previous experiences of all participants and innovative proposals applied to known restrictions. The initial step was the merging of mathematical modeling and information management through the Web. Paper forms to acquire patient's information were substituted by digital ones. In this way, all data could rapidly be accessed and available online. TB experts defined the data inputs and were responsible to validate the system and its outputs. Data quality methods were used to guarantee the accuracy of records, avoiding uncertain information and improving the value of the final result. Finally, portability was a mandatory requirement in order to guarantee the use of the system in different regions of the Country. The use of symptomatic information to feed a neural network model in order to build the decision support system would guarantee a reliable proposal, constructed with low cost resources.

2. Challenges in TB diagnosis

The increasing number of TB cases encourages the elaboration of modern, competent and economical feasible diagnosis methods. The diagnosis of SNPT is still a challenge

* José Manoel de Seixas[1], Felipe F. Grael[1], Rodrigo C. Torres[1], Fernando G. Ferreira[1], Andressa S. Gomes[1], José Márcio Faier[1], Jose Roberto Lapa e Silva[2], Fernanda C. de Q Mello[2], Afrânio Kritski[2] and João Baptista de Oliveira e Souza Filho[3]
[1]*Signal Processing Lab, COPPE/POLI, Federal University of Rio de Janeiro, Brazil*
[2]*Faculty of Medicine, Federal University of Rio de Janeiro, Brazil*
[3]*Federal Center of Technological Education Celso Suckow da Fonseca, Brazil*

although the availability of effective and suitable diagnosis tests. For the last 15 years, global tuberculosis (TB) control efforts have led to impressive results. However, despite these important achievements, the absolute numbers of TB cases continue to rise. In 2008, 9.4 million TB cases were estimated and 1.7 million individuals die globally (WHO, 2009). Highly effective and widely accessible diagnostics, drugs and vaccines are needed to address this problematic situation. Beside that, however, there are technical and structural challenges that impede optimal detection and treatment of all forms of TB cases in TB control programmes in different countries (Marais, 2010). TB control in most endemic countries relies heavily upon direct sputum smear microscopy, as this is most often the only simple test that can be used below reference laboratory level. Currently, however, only about 60% of all infectious TB cases are being detected with this test, and a proportion of those detected (i.e. listed in the laboratory registers as having at least one positive smear) do not come back to the clinic after submitting the first specimen, so do not receive appropriate treatment (Guillerm, 2006). Since direct smear microscopy is less sensitive in HIV-associated TB, further testing using complex technologies in sophisticated laboratories has been evaluated to reliably diagnose HIV/TB, in SNPT cases (Perkins, 2007). New simple inexpensive tools are needed to identify the various forms of TB (including drug-resistant and HIV-associated TB) at the lower levels of health services. Current assessment of the present diagnostics pipeline suggests that new tools will soon be available (Pai, 2010).

Since 2007, WHO has endorsed the use of at least 10 new diagnostic tools (technologies or approaches) that, if used wisely, could facilitate considerably TB control, and very recently endorsed a new automated real-time nucleic acid amplification technology (NAAT) for rapid and simultaneous detection of TB and Rifampicin resistance (Xpert MTB/RIF system) that offers drastically new prospects for the diagnostics of active and drug-resistant TB. Nevertheless, in a recent survey, 16 high-burden TB countries evaluated, about 50% of them are using TB diagnostic tools recommended by WHO from 2007-2009, NTP managers reported diverse challenges to the implementation of new diagnostics, but no impact assessment of its introduction on TB control was carried out (Van Kampen, 2010).

Therefore, there is insufficient evidence available to determine which package of current and newly developed diagnostic tests would work best in a given set of circumstances, and there is as yet little guidance available to countries on what new diagnostic tools, or combinations of tools, should be implemented in particular epidemiological/health system settings, with high prevalence of SNPT cases, and at what level of the health service it should be done. Recently, to address this issue, an impact assessment framework (IAF) was proposed and endorsed by WHO (Mann, 2010).

The decision support system corresponds to an innovative approach for SNPT diagnosis using statistical models, which might be useful in guiding health care workers in estimating the risk of SNPT, optimizing the use of more expensive tests for TB diagnosis.

3. Neural networks

Among statistical models that have been used to assist medical procedures, one can enumerate Bayesian networks, multivariate logistic regression, neural networks (El-Solh et al., 1999) and classification trees (Mello, 2006). Artificial Neural Network (ANN) is a biological inspired intelligence model, capable to learn through examples and to generalize, i.e. to produce coherent results to data not explored during learning process (Haykin, 2008).

This technique is attractive to a large scale of problems belonging to different domains, since it does not presume any statistical assumption about data variables or the problem itself and explores non-linear relationships between data variables to produce effective models, even in complex applications with a small number of available events in the dataset (Bishop, 2007), as typically occurs in medical applications.
The neural network mathematical problem modeling may provide extremely efficient and helpful tool in distinctive areas, mainly in diagnosis, prognosis and therapy, especially if models are properly formulated and data that is used to feed the neural network has good quality, reliability and represent a certain reality. When formulated in a systematic way and implemented with qualified data, statistical models can be representative of the clinical problem under evaluation and could be useful for physicians in their clinical routine, as well as for public health policy administration (Castelo et at, 2004).
Neural networks are made of a basic processing element known as neuron. Neurons are interconnected though synaptic weights forming the network. Roughly speaking, in the network learning process, knowledge is extracted from data and stored in these synaptic weights (Haykin, 2008). Different neurons models, network architectures and algorithms for training are available (Theodoridis, 2008). Basically, with respect to the learning process, neural models may be classified in supervised and non-supervised methods. Supervised training models are applicable to problems for which a desirable output network value (target vector) is available for each incoming data, i.e. tasks as function approximation and data classification. Non-supervised models, otherwise, identify groups which share similar statistical characteristics (clustering procedure) or extract specific features from data.
In order to produce realistic and accurate models, during the neural network development, one should focus on the selection of the variables which are problem representative, especially in the case of small datasets, which usually lacks of enough statistics to a better problem description as well as possesses class imbalance problems (Hastie, 2009). In this case, the selection of the training algorithm, the choice of the parameters involved in the algorithm, the complexity of the network structure adopted, the mechanism selected to control training, when applicable, are critical to obtain models with good generalization properties (Sahiner, 2008). Medical application usually also involves different kinds of variables, which should be properly coded to do not impact on network learning.
The decision support system here discussed explores both supervised as non-supervised neural models. Based on variables collected from triage, which were chosen by expert physicians on TB diagnosis, a multi-layer perceptron network – MLP- (Haykin, 2008), which is a supervised model, identifies the probability of a patient has TB. Additionally, a self-organizing network inspired on ART-2 algorithm (Vassali, 2002) assigns the patient to one of the following risk groups: low, medium or high. These methods provide independent and complementary information about the patient that is useful to support the decision taking process. In the sequence, these methods will be discussed in more details.

3.1 MLP neural network

The feedforward MLP architecture has neurons distributed into layers. The neuron model consists on a weighted sum of its inputs and applies this summing value to a non-linear

function, usually hyperbolic tangent, which is referred as the activation function (Haykin, 2008). Each layer has a defined number of neurons with forward connections to neurons that belong to the next layer. Usually, due to universal approximation theorem (Haykin, 2008), just three layers (input, hidden and output) are used in most of the problems. An arbitrary three-layer feedforward MLP network is shown in figure 1.

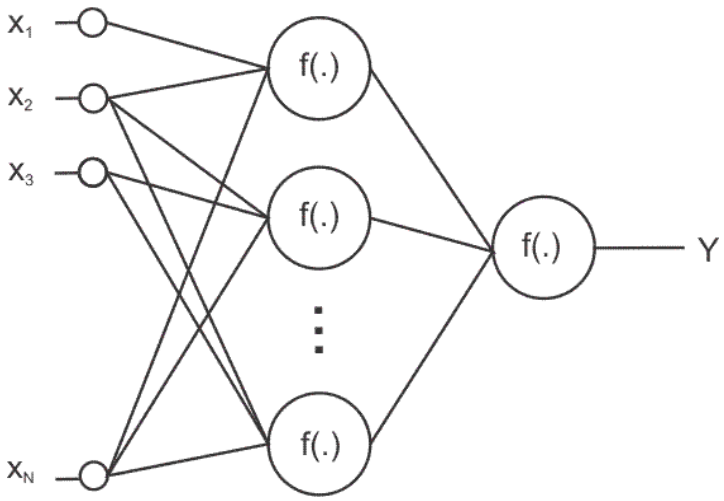


Fig. 1. Three-layer feedforward MLP network.

Considering three-layer feedforward networks, the number of hidden neurons has to be established, since the input nodes and output neurons are defined according to the quantity of input variables and to the desired network response, respectively. This choice should be according to the parsimonious principle (Medeiros, 2006), i.e. must follow a compromise between minimal complexity (few neurons) and maximum efficiency, otherwise the network generalization may be poor, especially for small datasets (Sahiner, 2008).

Another factor that may also jeopardize network generalization is the overtraining, which must be avoided through an appropriate training control mechanism control. A common adopted procedure is the early stop (Haykin, 2008), which consists on interrupting the training when the generalization error starts to increase. This procedure demands the construction of two disjoint sets: training and test, the first used during network learning phase and the last to evaluate generalization error. In applications with statistical restrictions, as the TB diagnosis modeling, training and test sets must be carefully chosen; otherwise learning may not occur properly. An inappropriate split may result in poor problem representation for training or non-realistic evaluation of network generalization, both prejudicial to model performance.

3.2 Self-organized ART-2 inspired network

This neural network explores a competitive learning algorithm which identifies groups of events that share similar statistical characteristics (Vassali, 2002). Each neuron responds to

a one identified group and defines a hypersphere in the input data space. The algorithm automatically determines neurons number and network connections needed to enclosure the data. Groups are described in terms of their center coordinates and radius, the last commonly referred as vigilance radius. Typically, all neurons share a same vigilance radius. In figure 2, clusters identified for arbitrary data by ART-2 inspired network are illustrated.
During network learning process, events should be presented to the network in a random order. Given an input event, the Euclidean distance from this event to the already identified group centers is determined. If this distance is lower than the vigilance radius, it means the event is inside at least one neuron coverage space. In this case, the neuron which shows lower distance (higher similarity) is declared the winner and has its center coordinates adjusted itself. If the event does not belong to any group, a new neuron (group) is created having the own event as its center. This process is stopped when: (i) no more neurons are created during training process or (ii) the center coordinates vary below a threshold between two consecutive iterations.

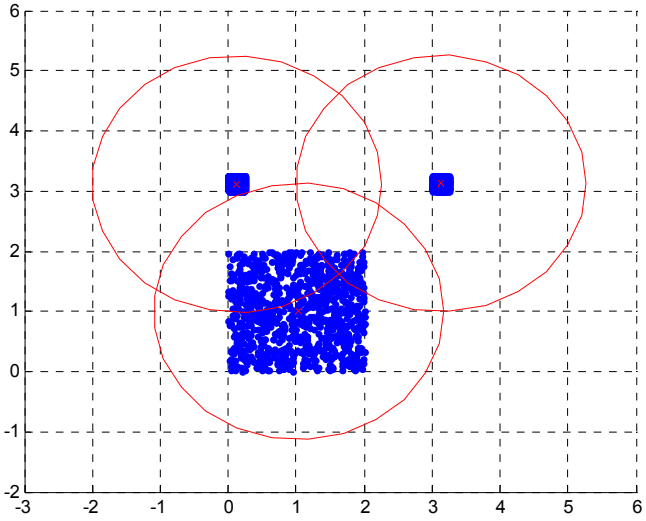


Fig. 2. Data clusters identified by ART-2 inspired network.

3.3 Pilot neural network models

Our feasibility studies to produce and apply neural models in supporting TB diagnosis started considering patients from the Clementino Fraga Filho University Hospital. In order to determine the set of symptoms and characteristics that would indicate the disease, one hundred and thirty-six patients agreed to participate. They were referred to the University Hospital from March, 2001 to September, 2002, with clinical-radiological suspicion of SNPT. The input data set for the neural networks model corresponds to information from clinical interview integrated to demographic and risk factors typically associated with tuberculosis diagnosis. All patients that took part in our research project were under suspicion of active pulmonary tuberculosis, presenting smear negative

results. Forty three per cent of these patients actually showed to have active TB. Initially, twenty-six clinical variables were considered: age, coughs, spit, sweat, fever, weight loss, chest pain, shiver, dyspnea, diabetes, alcoholism, and others. Later, the data set was described with twelve and eight clinical variables, selected by experts on tuberculosis research.

The dichotomy variables were codified as -1 and 1, representing the absence and presence of a symptom, respectively. Qualitative variables with three categories were implemented as -1 (lack of an indication), 1 (presence of the symptom) and 0 (ignored). Hyperbolic tangent was used as the activation function of all neurons from both hidden and output layers. All networks were trained using RPROP algorithm (Riedmiller, 1993) and the mean square error (MSE) between target vectors and network outputs was used as the objective function (Haykin, 2008). Networks having from 1 to 15 hidden neurons were produced. Due to local minima problem (Haykin, 2008), each network was trained using fifty different initialization parameters, being selected the training which showed high accuracy. Early stop was used as the learning mechanism control. Two approaches were considered to define the training and test sets used in early stop procedure. The first considered random splits of data. The second explored the partition of data on three clusters identified through the ART-2 inspired network. Seventy-five percent of the events belonging to each identified group formed the training set as the twenty-five remaining ones the test set.

The performance evaluation of the trained networks was based on the correct classification rate (accuracy over all patients), sensitivity and specificity. The best performance was obtained by the model having fifteen neurons and twelve variables (age, cough, fever, hemoptysis, anorexic, loss of weight, AIDS, night sweat, dyspnea, smokes actually, extra-pulmonary TB and ward hospital admission), which achieved 100% of sensitivity and 80% of specificity. This model was based on training and test sets selected by the clustering procedure. For the risk group identification, the vigilance radius of ART-2 inspired network was adjusted to produce three clusters. These groups were ranked in low, medium and high risk according their TB prevalence. The most frequent symptoms verified for each risk group were validated by TB experts.

4. The NeuralTB system

The NeuralTB, a decision support system for smear negative pulmonary tuberculosis, could be used by health care workers to sustain the diagnosis of SNPT under routine conditions in the hospitals and primary health care units. The purpose of using the system is not to replace physicians or other heath care workers, but to support them in decision taking. The proposed model suggests that mathematical modeling for classifying SNPT cases could be a useful tool for optimizing the utilization of expensive tests and to avoid costs of unnecessary anti-PT treatment, as may also permit an earlier diagnosis. The diagnosis corresponds to an ongoing process that requires accurate investigation and, therefore, the system output should be analyzed together with clinical interviewing, physical examination, and evaluation of laboratory results. One main concern for system design was to be suitable for areas of limited resources. Therefore, the requirements of low cost, easy access, and user-friendliness were considered. Since the disease in question is geographically spread among different places, the system is implemented using the Web technology,

guaranteeing its access. To places where the Internet is not available, the system can be easily installed in a portable computer. NeuralTB makes use of open source software, which makes the proposed system free of licensing costs, which comes to an inexpensive tool to support the identification of the various forms of TB and to be employed in different health services.

4.1 Main goals

The first step to develop the system to support the diagnosis of SNPT was to define its requirements. The NeuralTB system operates over symptoms and social characteristics of patients, which are used to feed the neural network model. SNPT specialists defined a set of symptoms and characteristics that would be presented in the first anamnesis procedure to be useful in a triage sector, usually the nurse attendant. The system also supports the registration of clinical procedures and laboratory exams to help physicians to start or not the anti-TB treatment and to make regimen changes furthermore, when indicated. The objective is to support the whole treatment from the arrival of a patient supposed to have TB, exams, drugs monitoring, till the medical discharge.

Electronic engineers designed, implemented, and trained neural networks, so that the system indicates the probability that a patient have SNPT or not and the risk group (low, medium or high) that this patient would belong to. The neural models were developed using data from patients with different TB modalities and associated co-morbidities. Software experts specified and developed the Web system to register the input information, execute the neural network processing, store the result, monitor the patients' data, and manage data files. The whole process is integrated with features that analyze, standardize, associate, clean and extract information according to data quality criteria: accuracy, completeness, updating, interpretability, security, access, privacy, consistency, etc. Professionals from the medical area supported the whole development process of the system design, validating each step to guarantee that the resulting system would achieve the predefined goals.

Another important requirement was to allow the use of the system in wireless devices. The software could be easily installed in a netbook allowing mobility and patient's treatment independently of his/her location. This approach is extremely useful to avoid treatment dropout since health agents may reach a patient at his/her residence and guarantee the continuation of the therapy. The whole system development was based on open source concepts, avoiding costs on acquiring software during either the implementation phase or implanting, using, and maintaining the product. The web system is installed in a central server that is accessed through the Internet. Nevertheless, in case a hospital or health care unit does not have Internet access, the computer can be connected to the main server through the 3rd generation of standards for mobile telecommunications services (3G). At last, in case the location does not have a good quality of mobile network coverage, the system can be installed in a local server. The interface was designed to be intuitive and user friendly, as normal site in the Web, which would prevent exhaustive training or any kind of assistance. To use the system, the only requirement for the end-users is a browser that usually is provided together with the operational system of the computer. The system is platform independent, it means that it does not require specific technological issues.

The central database manages the data of each hospital or health care separately. A location will access only the information that comes from itself and all other records are protected. For overall system management, fast and efficient retrieval mechanisms are provided for an administrator who will have full data access. The central system allows concurrent access of several locations. The data stored are certificated through intelligent computing methods in order to confirm certain characteristics of the information and avoid that distortions are recorded. The database design also considered the use of data profiling, i.e., analytical techniques used to examine existing data for completeness and accuracy. Data profiling is the first step towards data quality and would benefit the investigation of the disease, supporting the definition of policies to control tuberculosis.

4.2 Hardware and software requirements

The NeuralTB Web System runs over the Apache HTTP Server for both UNIX and Windows operating systems. The system provides a shell executable of setup programs that automatically install a directory structure and respective files in the computer of the health care unit or hospital. The minimum hardware requirements are: PC computers with a USB driver for file transfer (in case of local version) or an Internet connection, and having 128 MB, or preferentially, 256 MB RAM.

The system operations were implemented as CGI (Common Gateway Interface) programs, using the C language. The Javascript language is used to write functions that are embedded in or included from HTML pages and interact with the Document Object Model (DOM) of the page to perform tasks not possible in HTML alone. The Cascading Style Sheets (CSS) language is used to style the web pages written in HTML and format the XML documents. In order to draw the risk group representation, the GD graphics library was used. GD is an open source code library for the dynamic creation of images, allowing programmers to easily generate PNG, JPEG, GIF (among other images formats) from many different programming languages.

The central repository was implemented using MySQL, an open source relational database management system (RDBMS) that uses Structured Query Language (SQL). The choice of these technologies also facilitates the portability of the system to diverse platforms.

4.3 The web forms

The Web system forms allow users to enter data that are sent to the main server for processing and storage. Electronic forms facilitate not only the validation of input data but also the information access through the search of different fields of a document. It is also easy to make backups of the entire set of records and it does not spoil as easy as a paper form.

The web forms were implemented according to the paper form format and to the order of the questions, which minimizes the impact of introducing a new technology in the hospital or health care units. Users fill out the forms using checkboxes, radio buttons or text fields. Therefore, the web forms offer several advantages when compared to paper forms. They provide guidance, it means, the interface presents all possible choices and the user has just to select one. As an example, the marital status, the sex, the state where the patient lives, among other questions. Another type of guidance refers to conditional queries, for example,

the date of a previous TB treatment will be asked to the patient only if the previous answer to the question "previous TB treatment" is yes.

Another benefit of electronic forms is that they avoid submitting wrong data. Invalid entries are not allowed and an alert immediately pops up to notify the user that is filling the form. As an example, the user cannot insert a date where the month is greater than 12 or type letters of the alphabet to inform a zip code. Therefore, the system avoids typing errors. The forms were implemented to avoid submitting it if any question is not answered. The system automatically calculates one field according to the values of previous ones. As an example, the body mass index is determined after providing the weight and the height of the patient.

Together with the patient's data, the system records the date of the triage and name of the healthcare assistant or nurse that filled out the form. The information is retrieved from the login of the Web system and from the operating system. In case the patient does not know the answer of a question, there is an option to inform that this information is ignored. After saving the form, any further modification is completed together with additional information, such as the name of the user that modified the data, the reason for the change, the date when the alteration was performed and the previous value.

Figure 3 illustrates the anamnesis web form, where the patients' data are inserted. The forms were implemented to support the insertion of data related to the anamnesis interview, medical consultation, medical monitoring, exams and cost (that is split into two forms). The system informs in case a questionnaire was not properly filled out. The red color indicates that the form is complete while the black color points out forms that should be concluded.

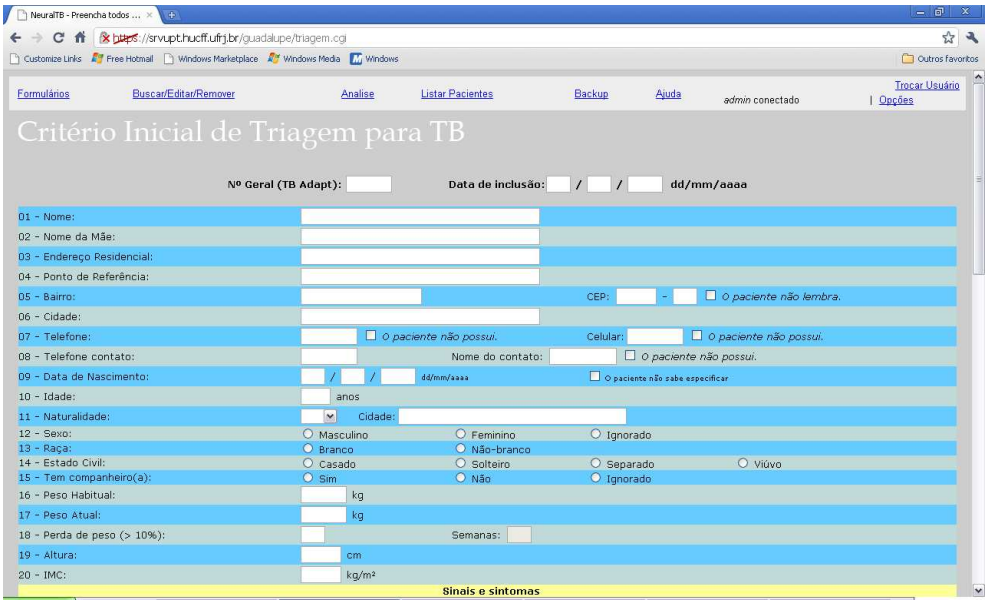


Fig. 3. Anamnesis web form. Fields are in Portuguese.

The system also provides several functions, such as, the search for a specific patient and the editing of patient data. The removal of a patient, for security reasons, corresponds to the absence of the patient data in the list but all the data continues to be stored in the database. The function analysis provides the percentage of a certain symptom or characteristic that can be combined by logical operators. It is also possible to list the name of all patients and then access the corresponding data. The user can export all records to an external program that would help building statistical graphics and can make a local backup. At last, the system provides a help that presents some explanation about how to fill out the forms.

4.4 Representing and assuring the data

Each item presented in the input data form is associated with the knowledge required to perform the anamnesis interview. Even facilitating the patients' data access, their analyses and interpretation is a laborious task and, in addition, records may contain redundant or incomplete information. Therefore, the plan was to represent the knowledge, patients' data, and other related details in a proper way. The data representation format is an important aspect that was considered to efficiently manage the whole information. Markup languages, as XML, can be used to describe knowledge structures and to support institutional memory development (Rabarijaona, 2000, Cook, 2000). XML may provide a standard structure to communicate and interchange data and knowledge among diverse systems. The language allows the creation of multiple visions of the same item and also provides an easy mechanism to capture, store, present and recover information. Considering these benefits, a XML-based approach was developed to describe the different types of knowledge and information manipulated during the whole process of the SNPT diagnosis.

There are three stages where data had to be properly represented: during the anamnesis interview, for describing the patient data, and to extract statistical information within research activities. TB specialists warned that risk factors, the questions made to the patients, and relationships among the stored records may vary according to locations or other factors, such as multidrug resistance (MDR), which is one of the main causes of ineffective treatment of new TB cases. Therefore, the use of XML facilitates the maintenance of the knowledge represented in the three stages. The tags identify the current data and new tags can be easily defined. The language also allows the definition of associations among diverse types of information.

In order to assure the compatibility between the data structure and the system functionalities, the Neural-TB interface and operations were conceived and designed in a way to guarantee its correct execution independently on both the way information is organized and the kind of records that are manipulated. This requirement is achieved by creating the interface with the system operations in the moment the application is executed. The interface reads the XML and presents all commands associated with the tags. So, in case one record type is excluded, the system will do not perform any operation related to this information. On the other hand, if a new record type is included, it is mandatory to define both the tag that identifies the data and the corresponding operation. This approach allows the execution of the system for the different versions of the stored knowledge related to the three stages of the process.

Another advantage of using XML is that it facilitates the integration among data that comes from different health care units and hospitals. Markup languages make easy the combination of heterogeneous records. The use of XML also allows uniform systems

interoperability and offers efficient mechanisms for information recovery. The interface between the knowledge and the neural networks model is also defined through a XML file. This archive describes the name of the application, the parameters used by the neural network, and the output. This approach facilitates when users want to execute a different neural network system or update the network parameters.
Data Quality refers to the adequacy of information to the needs, it means, the data have to be correct, do not duplicated, complete, reliable, consistent, standardized, updated, accessible, understandable to all category of users, such as patients, nurses, physicians, researches, etc. Increasing the quality of the data, the quality of the results produced by all processes that depends on this data is also enhanced.
The data quality monitoring is performed online to avoid error propagation and offline to improve the analysis of the data quality. The database is certified since all data is validated and corrected. This study can also come out with the measurement of the data quality, determining the magnitude of distortions and, therefore, identifying faults in the process that can be later corrected.

4.5 The pilot system

In order to demonstrate the feasibility of the proposal, a pilot system was designed. The main objective was to verify whether all requirements were accomplished and all necessities mapped. The chosen place was the Augusto Amaral Peixoto Policlinic, situated in Guadalupe neighborhood, Rio de Janeiro City, because the acquisition of patients' data had already started there using paper forms. The connection to the Internet in that site has a firewall that would prevent any link to the main server. In addition, the location does not have a good quality of mobile network coverage. In this case, our group had to install a proper server in the Policlinic. There, one room is dedicated to the triage and the second room is used to fill out the exam results and to the medical monitoring. A third room is used to the anamnesis interview. Since there were not other rooms available, the cost-effectiveness form could be filled out through a mobile device, such as a personal digital assistant (PDA), in any place of the Policlinic.
The infrastructure was organized to respect this arrangement. Three low-cost PCs with Windows operating system were installed in the three rooms. There was no need to install a special configuration since the only software requirement is the availability of a browser. The local server has a Linux operating system and a wireless network was established to provide connection to the PDA. Electrical secure systems were settled down there. Electric grounds were installed in each plug. High autonomy uninterruptible power source (UPS) were plugged to each computer where in case of lack of energy, the user's computer would run still one hour and the server computer would run still 6 hours. In order to guarantee that no data is lost, the server has redundant hard disk and power supply. The chosen operating system is Ubuntu since it is free, fast and secure. The access to the server is restricted to the administrators and protected with passwords. In case of faults, a spare PC, mouse, computer monitor, keyboard were available. The group made a set of studies to identify the robustness of the proposal and, consequently, the probability of failure of the architecture. Table 1 summarizes the performed analysis.
One copy of the system was installed in the local server while another copy was placed in the main server to allow the input of patient data previously registered in paper. For

security reasons, the paper forms could not leave the Policlinic in order to avoid lost of information. So, all paper forms were digitalized and a printed copy was sent to the typist. Another aspect that our team noticed was the fact that since the paper form had no guidance and no automatic error detection during the filling out procedure, the system version placed in the main server had to be modified in order to accept empty or not completely filled fields. Data conflicts had to be marked and later discussed with the physicians. The next step corresponds to the data validation by nursery technicians, who compared the original paper forms with the digital ones. A system to monitor the whole process of registering previously acquired data was also implemented. Therefore, one could check whether a paper form was digitalized, typed into the system and, finally, validated, guaranteeing the management of the information and its quality.

Component	Failure probability in the first year (%)	Recovery action	Recovery estimated time
PC (server)	0.017	Maintenance	2 days
PC (user)	16.0	Spare PC	30 minutes
Keyboard, mouse, monitor	34.35	Spare piece	15 minutes

Table 1. Failure probability, recovery action and recovery estimated time.

An additional feature had to be implemented since some data acquired during the medical monitoring was recorded into spreadsheet and data analysis proprietary software. The import functionality allowed the integration of data placed outside the system with the central database. The association is done using the identification number, patient's name and date of the interview.

Our group developed an automatic backup system to avoid data loss. All data is periodically sent to another server placed in the University Hospital and to external devices. Another web system gathers information of all backup files and monitors the number of patients and status of each form.

In order to support the activities in the health care unit, a control management system was also developed. It controls the versions of the forms, the data status and the access to the web system. Each time a new version of a form is created, the control system saves the previous version that are retrieved each time the user access data that was acquired through the earlier form. It is also necessary to keep track of the data origin, whether it was inserted directly to the NeuralTB system in a specific form or the information was written in paper and afterwards included in the system by the typist. In case of inconsistencies, a user can review the paper to check whether it was a typing mistake. Moreover, all data that was previously written in paper have to be verified after its insertion in the system. Without this validation, the data cannot be used for analysis. Therefore, the control system also administrates the two project repositories: one that contains data from the health care unit and other which data was entered by the typist.

The system access control is based on the users groups and their privileges, the system functionalities and the data itself. Within the system, there are three categories of users: administrator, physician and attendant. Administrators can insert new users, modify users

attributes and perform several actions related to the data files and system installation. Attendants may include a patient, symptoms, and edit registered data. Physicians is the user category that have all rights related to the patient's data and are the only ones that can see the artificial neural networks output, it means, whether the patient has SNPT or not and the risk group that the patient belongs.

4.6 System scalability

The pilot system demonstrated that each place can have its own requirements and the automation of routine cannot bring difficulties to the professionals who work in the hospital and health care units. Our group learned that is more efficient to adapt the software to each place than require each location to adapt to the system. So, in order to respect this requirement, an environment was build where different forms are defined and they can be combined to deliver a specific system for a certain hospital or health care unit. In this second part of the development, forms were implemented with Python language, an interpreted, interactive, object-oriented and extensible programming language. In order to merge the forms, our group used an open source web application framework, which follows the model-view-controller architectural pattern, thus facilitating the creation of database-driven web systems. The patient's data of each place is organized in local databases that use SQLite relational database management system. The source code for SQLite is in the public domain and implements most of the SQL standard.

Using 3G technology, it was not necessary to build a data transmission network among hospitals and health care units. Laboratories can connect to the system in order to insert the results of the exams, guaranteeing a high connectivity among different locations and the central database.

5. Impacts

The main objectives of the NeuralTB system are to increase the precision of an early TB diagnostic and to introduce a wireless triage associated with the neural networks. Therefore, this model avoids that patients are obliged to go to health care units, reducing queues and medical care delays. The benefits are extended to those patients who are unable to walk or have no resources to pay the transport. This technology will bring a cost-effective hospitalization approach since just patients with a high TB risk are placed in the respiratory isolation rooms or emergency health care units, resulting into a more efficient use of the government resources. So, for both patient and health care units, the proposal will bring a significant reduction of TB diagnosis and treatment costs. The research group also looks forward to enhance the TB control by reducing the number of new cases, with direct benefits to the whole population.

6. Conclusions

More than 4,000 patient forms are already registered in the main database. From this amount of records, 1,100 new patients were inserted directly through the system in the Policlinic of Guadalupe with the advantage that the system validates input data and alerts possible mistakes during the form filling process. Data quality techniques avoided recurrent problems during input data, such as duplicated records, typing errors, lack of information, non-standardized or not validated registers, incoherence data, inconsistencies, etc. In

addition, concerning information previously acquired, the data quality module assures the correctness and certification of the central database.

The repository gathers patient information from diverse regions and provides functionalities to easily handle a huge amount of data. The database contents can be migrated to other proprietary programs that provide specific features to analyze the data. Therefore, TB research groups can make use of the main database. Information privacy and secure data access are guaranteed by the system that encompasses functions to manage users and their permissions. Consequently, the system can keep track of modifications and register when and who carried out any data adjustment.

The use of markup language facilitates changes in the digital forms, which contributes to the continuously enhancement of not only the forms but also the whole process. Laboratories can connect to the system to directly insert exam results in the system. The complete patient treatment can be monitored wherever and the mobility allows home care by transmitting the data to the central repository. The system can incorporate specificities according to certain locations or contexts. The use of open source code assures the low cost of the software, which facilitates its installation and use in different environments. As a result, the solution corresponds to an economical feasible diagnosis method suitable to face the increasing number of TB cases

The decision support system can be considered as another element that helps physicians on the smear negative pulmonary tuberculosis diagnosis. The neural network model with fifteen neurons and twelve variables (age, cough, fever, hemoptysis, anorexic, loss of weight, AIDS, night sweat, dyspnea, smokes actually, extra-pulmonary TB and ward hospital admission) achieved 100% of sensitivity and 80% of specificity.

To incorporate this diagnostic test in the public systems, further investigation using the IAF proposed by Mann et al (2010) could be carried out in order to present important evidence on the costs and other resources required for the decision support system in the SNPT diagnosis implementation and optimized scale-up, along with the effects on patients and their clinical management, and TB transmission patterns in a variety of epidemiological settings. Additionally, through the policy transfer analysis, evidence about the processes that facilitates innovation uptake and policy transfer should be provided.

7. References

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*, Springer, ISBN 978-0387310732, New York, USA.

Castelo A., Kritski A.L., Werneck A., Lemos A.C., Ruffino Netto A., et al. (2004). Diretrizes Brasileiras para Tuberculose. *Jornal Brasileiro de Pneumologia*, Vol.30, suppl.1, (June 2004), pp. 1-86, ISSN1806-3713

El-Solh, A.A., Hsiao, C.-B., Goodnough, S., Serghani, J., Grant, B.J.B. (1999). Predicting Active Pulmonary Tuberculosis using an Artificial Neural Network. *Chest*, Vol.116, No.4, pp.968–973, ISSN: 0012-3692

Guillerm, M. (October 2006). Tuberculosis Diagnosis and Drug Sensitivity Testing: an Overview of the Current Diagnostics Pipeline. Campaign for Access to Essential Medicines. MSF, Paris.

Hastie T., Tibshirani R., Friedman J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, ISBN 978-0387952840, New York, USA.

Haykin S. (2008). *Neural Networks and Learning Machines*, Prentice Hall, ISBN 978-0131471399, New Jersey, USA.

Mann G, Squire S. B., Bissell K., Elisee v P., Du Toit E., Hesseling A., et al. (2010). Beyond accuracy: creating a comprehensive evidence base for TB diagnostic tools. *Int J Tuberc Lung Dis*, Vol.14, No.12, pp.1518–1524, ISSN 1027-3719.

Marais B. J., Raviglione M. C., Donald P.R., Harries A.D., Kritski A.L., et al. (2010). Scale-up of services and research priorities for diagnosis, management, and control of tuberculosis: a call to action. *The Lancet*, Vol. 375, No.9732, pp. 2179 – 2191, Published Online, DOI:10.1016/S0140-6736(10)60554-5

Medeiros M. C., Terävirta T., Rech G. (2006). Building Neural Network Models for Time Series: A Statistical Approach, *Journal of Forecasting*, Vol.25, No.1, pp.49-75, ISSN: 0277-6693.

Mello F. C., Bastos L. G., Soares S. L., Rezende V. M., Conde M.B., Chaisson R.E., Kritski A. L., Ruffino-Netto A., Werneck G. L. (2006). Predicting Smear Negative Pulmonary Tuberculosis with Classification Trees and Logistic Regression: a cross-sectional Study. *BMC Public Health*, Vo.6, No.43, Published Online 2006, DOI:10.1186/1471-2458-6-43

Pai M., Minion J., Steingart K., Ramsay A. (2010). New and improved tuberculosis diagnostics: evidence, policy, practice, and impact. *Curr Opin Pulm Med.*, Vol.16, No.3, pp.271-284, ISSN 1070-5287.

Perkins M. D., Kritski A. L. (2002). Perspectives. Diagnostic Testing in the Control of Tuberculosis. *Bulletin of the WHO*, Vol.80, No.6, pp.512-513. ISSN 0042-9686

Cook J. (2000). XML Sets Stage for Efficient Knowledge Management, *IT professional*, Vol.2, No.3, pp.55-57, ISSN: 1520-9202

Rabarijaona A., Dieng R., Olivier C., Quaddari R. (2000). Building and Searching an XML-Based Corporate Memory, *IEEE Intelligent Systems*, Vol.15, No.3 pp.56-63. ISSN 0884-8173

Riedmiller M., Braun H. (1993). A Direct Adaptive Method for Faster Backpropagation Learning: the RPROP Algorithm, *Proceedings of the IEEE Conference on Neural Networks*, pp.586-591, ISBN 0-7803-0999-5, San Francisco, California, USA, 28 Mar ch 28 - April 01, 1993

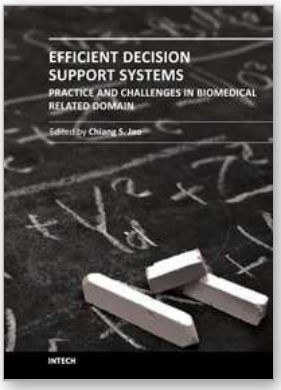
Sahiner B., Chan H.-P., Hadjiiski L. (2008). Classifier performance estimation under the constraint of a finite sample size: Resampling schemes applied to neural network classifiers, *Neural Networks*, Vol.21, No2-3, pp.476-483, ISSN 0893-6080

Theodoridis S., Koutroumbas K. (2008). *Pattern Recognition*, Academic Press, ISBN 978-0126858754, Amsterdam, Netherlands

van Kampen S. C., Ramsay A. R., Anthony R. M., Klatser P.R. (2010). Retooling national TB control programmes (NTPs) with new diagnostics: the NTP perspective. *PLoS One.* Vol.5, No.7, p.e11649, ISSN 1932-6203

Vassali M.R., Seixas J.M., Calôba L.P. (2002). A Neural Particle Discriminator based on a Modified ART Architecture, *Proceedings of the IEEE International Symposium on Circuits and Systems*, Vol.II, pp.121-124, ISBN 0-7803-7448-7, Phoenix-Scottsdale, Arizona, USA.

WHO – World Health Organization. (2009). *Global tuberculosis control: a short update to the 2009 report*. WHO/HTM/TB/2009.426. ISBN 978 92 4 159886, Geneva.



Efficient Decision Support Systems - Practice and Challenges in Biomedical Related Domain

Edited by Prof. Chiang Jao

ISBN 978-953-307-258-6
Hard cover, 328 pages
Publisher InTeh
Published online 06, September, 2011
Published in print edition September, 2011

This series is directed to diverse managerial professionals who are leading the transformation of individual domains by using expert information and domain knowledge to drive decision support systems (DSSs). The series offers a broad range of subjects addressed in specific areas such as health care, business management, banking, agriculture, environmental improvement, natural resource and spatial management, aviation administration, and hybrid applications of information technology aimed to interdisciplinary issues. This book series is composed of three volumes: Volume 1 consists of general concepts and methodology of DSSs; Volume 2 consists of applications of DSSs in the biomedical domain; Volume 3 consists of hybrid applications of DSSs in multidisciplinary domains. The book is shaped decision support strategies in the new infrastructure that assists the readers in full use of the creative technology to manipulate input data and to transform information into useful decisions for decision makers.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821