Automated diagnosis and disease characterization using neural network analysis

Carlo Moneta, Giancarlo Parodi, Stefano Rovetta, Rodolfo Zunino
Department of Biophysical and Electronic Engineering (DIBE)
University of Genoa - Italy

Abstract — **A neural network approach is used to analyze and diagnose a rather new and uncommon disease, Lyme borreliosis. In order to fully exploit the method's generalizing power, a significance analysis split the set of inputs of a trained network into two classes (important and unimportant); the results of this analysis lead to a new "structured" network, whose topology and architecture reflects the estimated relevance of symptoms. The diagnostic performance thus obtained shows a dramatic improvement.**

I. INTRODUCTION

A. Motivations and baseline

Medical diagnosis problems represent a classical and sound testbed for Pattern Analysis techniques for many reasons. First, medical data are often easily translated into "patterns"; then, they have been pre-processed by experts with a deep knowledge of the medical problem, thus bypassing unsafe, often improper processing by non-expert researchers. Finally, both knowledge engineers and medical researchers can benefit from an interdisciplinary cross-check; on one side, pattern analysis techniques can be tested at solving a "real world" problem, whereas, on the other hand, automated methods can be investigated to extract information from medical databases.

Neural networks (NN) represent an effective tool to perform this kind of tasks. They offer the possibility of being trained "by examples", i.e., learning a given data set. Moreover, with their remarkable generalization power, new data not included in the training set can be correctly interpreted; this is a result of the inner (implicit) representation of the database, learned during training.

The capability of catching the underlying structure of a data set is an important and useful feature of a NN. On the other hand, this feature itself often requires expert knowledge to interpret results and to tune a specific classification system. Consequently, such methods generally address problems in which complete expert knowledge is available (e.g. [4]).

The problem tackled in this research is the diagnosis of the Lyme disease. It differs from the above ones substantially: the medical state-of-the-art does not provide yet an exhaustive diagnostic procedure. Many "standard" criteria have been proposed but, despite some peculiar symptoms, a fixed set of discriminant symptoms can not be identified; this is mainly due to the extreme variety of the possible manifestations. In addition, the actual characteristics are often disguising and lead to confusion with other pathologies. This means that not all recorded observations have the same importance, some being really symptoms for the disease, other secondary characteristics, and other being even unuseful data, recorded for the sake of completeness.

This problem statement imposes a twofold research goal. On one hand, we want the diagnostic system's performance to be improved. On the other hand, in order to obtain the best performance, it is necessary to assess the relevance of the different symptoms in patient descriptions.

The NN model used for these tasks is the standard feed-forward network, trained with the back-propagation algorithm. The symptom-relevance problem has been tackled by defining and evaluating the sensitivities of the NN outputs to input symptoms. Such analysis led to an importance-ranking ordering of the various medical observations. The results of this step always confirmed the expectations suggested by medical experts.

The huge number of the describing symptoms (84 network inputs) heavily complicates classification. The related increase of number of weights is an obstacle to proper generalization. Therefore, an unsupervised data-compression process treated a set of poorer-significance symptoms (42 symptoms, as ranked out by the previous analysis), shrinking them to a compressed representation of five coding values. This reduced the number of inputs from 84 to 47, i.e., 42 high-significance symptoms and five "coding" values.

Thanks to this information-coding step, diagnostic performance reached an average error rate around 6%. These figures, while reflecting the complexity of the actual clinical problem, represent a notable success as far as standard medical diagnosis rate is concerned. From a technical point of view, the result of the described research is a general methodology to handle complex classification problems.

Manuscript received August 1, 1992

0-7803-0720-8/92 $3.00

B. Clinical context

From a medical viewpoint, the study of Lyme disease through pattern analysis is complicated by two factors: the inadequacy of computer tools currently available for medical applications, and the difficulty inherent to the disease itself.

Attempts to implement computer-aided diagnostic processes using AI techniques are far away from technical significance [2]. Most of them are modeled on classical expert systems (rule-based reasoning programs); each of them features some variation on the basic model, anyway evidencing that none is completely satisfactory. Average error rates range from 10% to 25% or even more, depending on how general a system is intended to be (that is, on the number of diseases the system knows).

As far as the general computer-aided diagnosis scenario is concerned, this is not a very good situation. In the Lyme-disease case, additional difficulties arise specifically from the problem considered. Lyme disease was discovered and identified in 1977 (first cases studied in 1975, coming from Old Lyme, Connecticut), but some of its aspects (under different names and descriptions) had been known since the early XX century. This is due to the extreme variety of forms it can present. Some of its symptoms are very peculiar but also quite rare (e.g., the typical skin manifestation called Erythema Chronicum Migrans - ECM). Other symptoms are very confusing, and can mislead diagnosis: they range from arthritis to neurological involvement to cardiological complications. To make things even worse, only in a few cases the list of symptoms is clearly related to Lyme disease: in general, only a few of all peculiar characteristics are present, and they may span over a time interval ranging from months to years.

Lyme disease is caused by *Borrelia Burgdorferi*, a spirocaeta carried by different kinds of tick (*Ixodes Dammini*, *Ixodes Ricinus*). This means that the disease is present only in specific geographical areas, namely, where: 1 — there are animals parasited by ticks; 2 — ticks are parasited by *Borrelia Burgdorferi*; 3 — humans have the chance to come in *frequent* contact with ticks (statistically, it is not so easy to be infected). In summary, Lyme disease is not only difficult to recognize, but also relatively rare to find.

In the experiments, the database included 741 samples of patients; almost 200 of those were classified as Lyme-affected, hence the rest was a set of counterexamples. After a pre-processing of these data, we defined 84 describing fields for each subject, including both general information (e.g., age) and specific clinical recordings (e.g., presence of ECM).

C. Neural context

The basic topology of the NN has 84 input units, with continuous input values, and one hidden layer with 4-6 units. The output layer consists of 2 units, representing diagnostic categorizations suggested by physicians (ranging from surely ill to surely unaffected). These values are selected on an analogical scale, that could be coded by a single real number ranging, say, from —1 to +1. With two output values the NN can represent the same information in two ways, and reduce uncertainty by introducing redundancy. Therefore, one output unit represents the level for the target attribute "ill", whereas the other unit codes the attribute "unaffected".

TABLE I
LYME DISEASE

Caused by Borrelia Burgdorferi, brought by Ixodes Dammini, Ixodes Ricinus and other ticks
Found in specific geographic locations
Early phase (2 weeks since infection):
- ECM
- Non-specific symptoms (fever, weakness)
- Specific anticorpal response: IgM ≥ 256:1 (standard threshold)
Middle phase (1 to 6 months):
- Cardiac symptoms
- Neurological symptoms
Late phase (1 year or more):
- Arthritis, responding to antibiotic therapy
- Specific anticorpal response: IgG ≥ 256:1

The training algorithm used is called SuperSAB [5], and is an accelerated version of classical back-propagation. In SuperSAB, an individual learning step is assigned to each possible dimension in the weight space, and is adaptively modified during training. By combining this optimization technique with the implementation on RISC technology, the computation times required were limited to an acceptable amount.

It turns out that the number of available subjects (that is, training patterns) will not enable the NN to generalize properly during the training phase [1]. With a delicate (but often used) procedure, new training samples have been generated when required, based on the available data set with the addition of random noise [3].

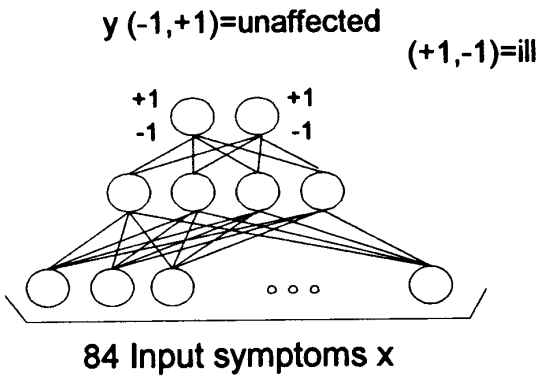

Fig. 1 — Diagnostic neural network

II. RESEARCH DESCRIPTION

A. Significance Analysis

Evaluating the relevance of each network input makes it possible to order the describing attributes (features). To this end, instead of applying statistical methods to the data set, an already trained network has been used. The NN training results in a group of weight matrices, one for every two consecutive layers. These weight sets express in a distributed form the information extracted from data during training; they can be used to estimate the relevance of each feature. Such an approach has the appealing characteristic to be wholly "network-oriented". In other words, we do not try to analyze the database with non-neural techniques, and then apply the results to neural methodologies; rather, we "ask the network its own opinion," so we can expect coherent results in each phase of the research.

Arranging input features according to their relevance has two purposes. First, providing experts with a numerical analysis can confirm their qualitative estimate and give a first cross-checking result. Second, poorer-relevance features can be identified, thus compiling two lists of attributes (important and unimportant). Those features belonging to the second class have a smaller information content and could be removed. Alternatively, some rearranging can be looked for, and a smaller feature set (less network inputs) would consequently improve performance.

Now we will go through a few mathematics. The ultimate goal is to evaluate the effect that the variation of a quantity, x, has on another quantity, y, (sensitivity analysis). This is usually done by calculating the derivative of the function $y(x)$ with respect to x. From this point of view, a trained NN expresses a nonlinear (vector) function of a (vector) variable, with the weights representing adjusted parameters. Therefore, the variation of the j-th network output caused by a variation of the i-th input can be measured by:

$$r_{i,j}(\mathbf{W},\xi) = \left.\frac{\partial x_j^{(3)}}{\partial x_i^{(1)}}\right|_{x_i^{(1)}=\xi} \tag{1}$$

where:

- $x_k^{(n)}$ is the activation value of the k-th unit in the n-th network layer;
- $\mathbf{W}$ is a vector containing all weights and biases for the network;
- ξ is the specific input value for which the calculation is made.

The analytical expression of (1) can be derived from the definition of neurons' activation functions. Let $f\left(net_i^{(n+1)}\right)$ be the activation function of the i-th unit of the n+1-th layer, applied to the weighted sum of inputs (coming from the n-th layer):

$$net_i^{(n+1)} = \sum_{k=1}^{N^{(n)}} w_{ki}^{(n)} x_k^{(n)} \tag{2}$$

Notice that the first subscript of weight terms indicates the "from-unit", the second indicates the "to-unit", and the superscript indicates the "from-layer". The derivative of the j-th output with respect to the i-th input can now be calculated for a three layer network.

$$\frac{\partial x_j^{(3)}}{\partial x_i^{(1)}} = \frac{\partial \mathrm{f}\left(net_j^{(3)}\right)}{\partial x_i^{(1)}} =$$

$$= \frac{\partial \mathrm{f}\left(net_j^{(3)}\right)}{\partial net_j^{(3)}} \frac{\partial net_j^{(3)}}{\partial x_i^{(1)}} =$$

substituting (2):

$$= \frac{\partial \mathrm{f}\left(net_j^{(3)}\right)}{\partial net_j^{(3)}} \frac{\partial \sum_{k=1}^{N^{(2)}} w_{kj}^{(2)} x_k^{(2)}}{\partial x_i^{(1)}} =$$

$$= \frac{\partial \mathrm{f}\left(net_j^{(3)}\right)}{\partial net_j^{(3)}} \sum_{k=1}^{N^{(2)}} w_{kj}^{(2)} \frac{\partial x_k^{(2)}}{\partial x_i^{(1)}} =$$

$$= \frac{\partial \mathrm{f}\left(net_j^{(3)}\right)}{\partial net_j^{(3)}} \sum_{k=1}^{N^{(2)}} \left[\frac{\partial \mathrm{f}\left(net_k^{(2)}\right)}{\partial net_k^{(2)}} w_{kj}^{(2)} \sum_{h=1}^{N^{(2)}} w_{hk}^{(1)} \frac{\partial x_h^{(1)}}{\partial x_k^{(1)}} \right] =$$

$$= \frac{\partial \mathrm{f}\left(net_j^{(3)}\right)}{\partial net_j^{(3)}} \sum_{k=1}^{N^{(2)}} \left[w_{ik}^{(1)} w_{kj}^{(2)} \frac{\partial \mathrm{f}\left(net_k^{(2)}\right)}{\partial net_k^{(2)}} \right]$$

Given a specified weight set $\mathbf{W}$ (i.e., when the NN has been trained), *each* training pattern provides an estimate r_{ij} of the relevance of the j-th symptom on the i-th output for a particular value ξ of the input. Therefore, this kind of sensitivity analysis leads, for each output unit, to as many estimates as the number of training patterns.

To generalize results, it is necessary to compute a statistical characterization these data. In the presented methodology, such a description is obtained by estimating average value and variance of r_{ij}:

$$R_{i,j}(W) = E_x\{r_{i,j}\}$$
$$\sigma^2{}_{i,j}(W) = E_x\{r_{i,j}^2\} - E_x{}^2\{r_{i,j}\}$$

By combining the values obtained for the two R_{ij}'s (with j=1 and 2 - two output units), one can get some hint about the relevance of the i-th symptom to the overall classification outcome. The techniques used in this interpretation process may range from purely statistical analysis to massive knowledge-based approaches; in our experiments, a simple comparison rule of the two values was used, without affecting the method's generality. However, the overall result is a rule expressing some significance evaluation.

As this is only a rough estimate, some reliability measure for inferred rules is also required. This can be accomplished by analyzing variances together with average values. For instance, low average and low variance will suggest marked irrelevance. By contrast, low average value with high variance will indicate some uncertainty, in which irrelevance is not ensured; in such cases, the reason must be likely searched for in a correlation of the considered symptom with other features.

The numerical analysis of Lyme disease data yielded a list of symptoms, in order of importance. By a suitable thresholding operation, the top and the bottom portions of the list were extracted to form two symptom groups for medical experts (12 unimportant, 10 important features). The "central" segment, containing not extreme cases, turned out to be relatively wide because of the sparsity of information through the features; the involved symptoms have not been taken into account.

The two lists of symptoms have been of great usefulness for medical experts: first, results confirmed qualitative expectations and supported experts with numerical evidence; second, the analysis also helped physicians in pointing out irrelevant informations they could avoid to collect.

TABLE II
ESTIMATING THE RELEVANCE OF SYMPTOMS

Values of the two R's	Remarks about the symptom
Low	Low relevance
Only one high	Characterizing but not discriminating
Both high, same sign	If σ^2 's are high, there is strong correlation with other symptoms. Otherwise, there must be something wrong (inconsistent result)
Both high, different sign	Relevant feature
Values of the two σ^2's	**Remarks about the symptom**
Low	Relevance observation is reliable
One is high	Probable correlation
Both high	Relevance observation is not reliable

B. Improving diagnostic performance

The complexity of training demands some reduction of data dimensionality. This is necessary to ensure proper generalization from the limited number of examples; moreover, it would keep CPU time within acceptable limits, thus making many repeated trials feasible.

Reducing data dimensionality requires to identify a higher number of features for pre-processing. The previous analysis pointed out the *most* irrelevant features. By lowering the threshold on the histograms of average and variance, one can identify a set of 42 symptoms that appear "*more*" irrelevant than the others, although not completely unuseful. Therefore, a cut in the feature list is no longer appropriate, whereas a data-encoding step might reduce the inputs' number without losing sparse, but still useful information.

Those 42 features have been compressed into 5 coding values, using an auxiliary network trained by auto-associative mapping.

In this schema, the 42 inputs act as input and output patterns at the same time. The training goal for this NN is mapping an input pattern onto itself through a 5-unit bottleneck (Fig. 2). The activation values of the five bottleneck units encode the compressed output. It is worth stressing that no supervising information is provided in this phase. As no target value other than the input itself is given during training, the method generality is fully preserved.

The coded output can then join the remaining features, building up a new vector with lower dimensionality (42 descriptors + 5 coding output values = 47 features, instead of 84). This vector can then be used to feed the input layer of a conventional NN for normal (supervised) training. This procedure is summarized in Fig. 3, where the two cascaded NN's are integrated in a single structured network.

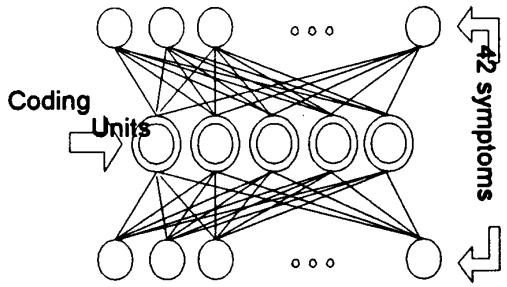

Fig. 2 — The coding neural structure.

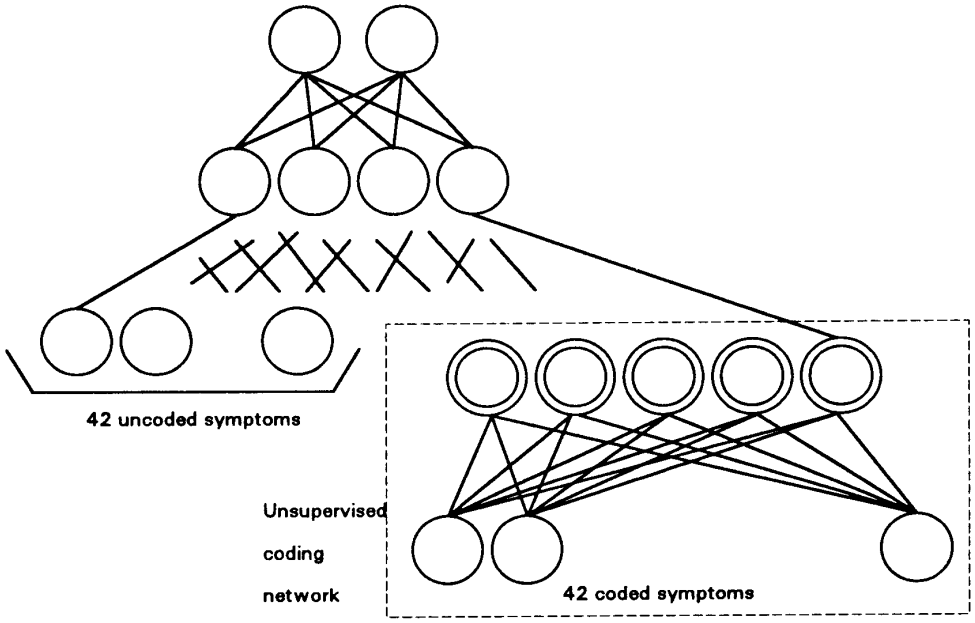

Fig. 3 — The diagnostic network integrating the two neural structures.

III. RESULTS AND COMMENTS

A. Experimental results

The method has been tested by comparison with: 1) the classification performance of a conventional NN trained by the standard supervised back propagation algorithm (SuperSAB); 2) the classification performance of a supervised-coding approach, in which coding used supervised training with correct target values.

The target used the two-unit code described in section I, without half-tones (that is, all possible levels were compressed to [+1,—1] for unaffected and [—1,+1] for ill).

In each trial, a test set of about 10% of the total database has been kept disjointed from the training set, to have a reliable measure for the performance, thus expressed as error rate on new patterns.

Different trials were made changing the selection of training and test set. Test sets always featured the same proportions between the two classes (ill/non ill) as the original data set.

Table III summarizes the results. It can be seen that an error rate of 21% in the conventional training reduced to about 12% in the supervised coding and reached 6% in non-supervised coding, with an error rate near (most often equal) to 0% on ill subjects.

TABLE III
EXPERIMENTAL RESULTS

Network structure	Error rate	Error on each class	
		Ill	Non-ill
Standard	21%	27.3%	19.8%
Supervised coding	12%	15%	4%
Non-supervised coding	6.2%	0%	7.4%

B. Comments

This research showed that combining significance analysis with a coding process notably enhances the system's diagnostic performance. This holds not only for numerical ratings, but also from a clinical point of view. In other words, the clinical reliability of the results, together with the average error percentage, has been kept under observation with medical experts supervision, showing real improvements.

Unsupervised coding outperforms not only standard NN, but also the method's supervised-coding version. This can be explained by noting that auto-associative training provides a deeper "understanding" of the underlying structure of the data set. As shown in Table III, error distribution among the two classes is not constant. In a normal situation (standard NN), ill subjects feature the greater error percentage. As they are a minor part of the data set, patterns must be artificially generated to equilibrate the training set and to avoid Bayesian polarization of training. Bayesian polarization occurs when the relative frequencies

of classes are learnt, independently from any other pattern characteristics. By applying unsupervised coding with a suitable attribute selection, this undesired distribution is minimized and even inverted: ill subjects may eventually feature a lower error rating than non-ill subjects (in this case, reaching 0%). This shows that compression does not remove relevant information, but, on the contrary, such information is somewhat enhanced by the reduction of data dimensionality.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Prof. Rovetta and the whole staff at the Centro Reumatologico Ist. Bruzzone — USL XVI — Dept. of Internal Medicine of Genoa University — Italian Working Group on Lyme disease, for providing both data and valuable assistance.

REFERENCES

[1] A.C. Steere, S.E. Malawista, D.R.Snydman, "Lyme arthritis: an epidemic of oligoarticular arthritis in children and adults in three Connecticut communities," *Arthr Reum*, 20:7, 1977

[2] W. Burgdorfer, A.G. Barbour, S.F. Hayes, J.L. Benach, E. Grunwaldt, J.P. Davis, "Lyme disease — A tick-borne spirochetosis?", *Science* 216:1317, 1982

[3] E.C. Baum, D. Haussler, "What size net gives valid generalization?", *Neural Computation*, No. 1, 1989, pp.151-160

[4] H. Bernelot Moens, J.K. van der Korst, "Computer-assisted diagnosis of rheumatic disorders", *Seminars in Arthritis and Rheumatism*, vol.21, No.3, Dec 1991, pp.156-169

[5] L. Holmstrom, P. Koistinen, "Using additive noise in back-propagation training," *IEEE trans. on Neural Networks*, vol.3, No.1, Jan 1992, pp.24-38

[6] R. Poli, S. Cagnoni, R. Livi, G. Coppini, G. Valli, "A neural network expert system for diagnosis and treating hypertension," *Computer*, march 1991, pp.64-71

[7] T. Tollenaere, "SuperSAB: fast adaptive back propagation with good scaling properties," *Neural Networks*, vol.3, 1990, pp. 561-573