Proceedings of the 8th IEEE Symposium on Computer-Based Medical Systems
Lubbock, Texas, USA, June 1995, pp. 57-64.

Validation of a Large Medical Database

Guido Rovetta,° Patrizia Monteforte,° Gerolamo Bianchi,°
Stefano Rovetta,* and Rodolfo Zunino*

University of Genova, Italy
* Faculty of Engineering, DIBE ° Faculty of Medicine, DIMI

Abstract

Complex clinical problems involving huge experimental evidence require a preliminary validation of observed data. This may avoid biasing due to incorrect sampling and clarify the sample distribution by showing data-inherent regularities. The paper describes the application of unsupervised models of neural networks to the analysis of a very large set of clinical records for the study of Osteoporosis. The main result obtained lies in showing the overall uniformity of the data distribution, which indicates a correct, unbiased sampling of the considered population.

1. Introduction

The Ist. Bruzzone Rheumatological Center (Department of Internal Medicine, USL 3) has recorded a large database of epidemiological observations related to Osteoporosis. This work reports on some studies conducted on this database as a preliminary step, prior to the effective extraction of significant informations.

In the last years, data-analysis methods have become very sophisticated. However, the attention tends to be biased towards the developments of increasingly powerful algorithms. This involves that sound validation methods for the data should be improved accordingly. There seems to be a gap between the two requirements: the development of methods such as pattern recognition techniques (in the 60s-70s), knowledge-based systems, nonlinear and nonparametric statistical procedures, neural networks (80s-90s) has not caused a proportional attention to the goodness of the data properties. This work deals with the application of a non-standard validation method, using a neural network approach, to a critical and significant clinical problem.

2. Clinical context and problem statement

Osteoporosis is currently a major problem in medicine, for a variety of reasons. It is a cause of fractures and invalidity in elder people, with a notable social cost. Only global criteria can be applied. A pathological diminution in the bone mass can be assessed only with specific instrumental observations (Single photon absorptiometry, Dual energy X rays absorptiometry, Quantitative ultrasound [1]) and both value and trend in time should

1063-7125/95 $4.00 © 1995 IEEE

be analyzed. The diagnosis of Osteoporosis is a fuzzy problem, in that it is not described by a definite threshold or yes/no value. The range of normal values is assumed to be ±2 standard deviations from the average value, and is parametrized by sex, age and time of occurrence of menopause. This soft threshold is used to assess the probability of fractures, in conjunction with life habits and other factors.

A natural consequence of both the complexity and the relevance of the clinical problem is the huge amount of experimental observations usually collected. In other words, a great quantity of observed patients, with many parameters for each person, is required to approach the diagnosis [2]. This raises another crucial issue, concerning the significance of experimental evidence from a statistical perspective. More precisely, the prevalence of the clinical phenomenon requires that the distribution of available data (which help tuning diagnostic performance) be thoroughly validated.

From this viewpoint, this paper addresses advanced connectionist techniques for an extensive analysis of data distribution. Thus the overall research goal is not a simple improvement in classification accuracy, but rather to inspect data to reveal regularities and distribution features. The described research adopts unsupervised representation methods, which group experimental observations according to similarity criteria. The involved clustering process does not take into account a patient's actual clinical diagnosis, but tends to find out descriptions of "natural" patient groups reflecting data-intrinsic structures.

This determines a two-fold research objective: on one hand, analyzing the distribution of samples may help validate the data collection process by removing sampling peculiarities; for example, the presence of a huge group of patients belonging to an oversampled category might distort the environment representation and bias further data-analysis processes. On the other hand, a consistent grouping of observed data may lower the clinical problem's complexity by reducing its dimensionality. In this case, the underlying assumption is that a good clustering may satisfactorily represent the entire data collection by means of a *much smaller* set of prototypes (vocabulary). This representation process is very common in pattern-recognition applications, and is called Vector Quantization (VQ) [3].

A preliminary research step exploits classical neural models, such as Kohonen's Self-Organizing Feature Maps (SOMs) [4], which position a (fixed) number of reference vectors at significant locations in the data space. An important feature of this model is that the final network configuration is topologically consistent with the actual data distribution.

The use of SOMs, however, does not help one solve another important issue, namely, determining the proper number of prototypes to be used in the quantization process. A small prototype set might be ineffective for detailing all data-intrinsic structures, whereas too large a vocabulary might prove excessively detailed and computationally impractical. Therefore, we developed a novel technique to tackle the vocabulary-dimensioning problem. The "plastic" method adaptively adjusts the cardinality of the prototype set, and follows a statistical cross-validation approach to determine a network's generalization ability.

3. Clinical data

The database gathers about 12 years of observations, relating to 16 000 people coming from the area of Genoa, in the region of Liguria (Northwestern Italy). Since the observations cover an area which can be estimated to count about 800,000 people, we can assume that descriptions of 2% of the whole population has been recorded in the database (note, however, that only people aged over about 40 are subject to osteoporosis, so the percentage should be higher). Bone density is higher in males than in females, and decreases with age. Hence the typical observed person is a woman aged over 40; nevertheless, men are also present. 190 observed parameters report on many aspects of life, such as job, alimentary habits and intolerances, physical activity, all known or suspected to be related to osteoporosis. These recordings are unique for each patient; in addition, the SPA parameters are recorded more than once, to monitor their evolution in time.

The data base, recorded with the aid of a commercial software, has been pre-processed to help applying user-developed methods to data analysis. The conversion to numerical values of all continuous valued, yes/no and multiple-choices variables has yielded a record size of 274 numerical fields, excluding the SPA values. Each numerical record can be viewed as a vector in a 274-dimensional space. In principle, one would expect that the distribution of these vectors in this space be as uniform as possible for the data base to represent a statistically valid sample. Since the patient-sampling strategy depends on environmental conditions and is therefore most difficult to control, the validity condition for the database must be verified *a posteriori*. After this verification, if the distribution exhibits some natural groups, a link between cluster compositions and diagnostic outcomes can be searched for.

4. Neural methods

4.1. Self-Organizing Maps

In a preliminary research step, we used Kohonen's Self Organizing Maps to inspect the data distribution. This model positions a set of representative vectors (prototypes) in the data space by an iterative procedure. The training process aims at placing prototypes at positions in the data space such that their configuration spans a "good" representation of the whole data set. The quality of a vector configuration is measured by a *cost function*, typically Mean Square Error:

$$MSE = \frac{1}{N_p} \cdot \frac{1}{d} \sum_{p=1}^{N_p} \sum_{j=1}^{d} \left(x_p^{(j)} - w_{B(p)}^{(j)} \right)^2 \tag{1}$$

where: N_p is the total number of samples; d is the data-space dimensionality; $\mathbf{w}_{B(p)}$ is the prototype that is closest to the p-th sample (i.e., it is the prototype lying at the smallest Euclidean distance from the sample).

In Kohonen's SOMs, vectors are arranged in a fixed topological structure determining each neuron's "neighborhood." In the research presented here, a two-dimensional neighborhood lattice including 8x8 neurons has been used as a default. The method implements a *Winner-Take-All* (*WTA*) strategy, in which, for each training sample, the best-matching neuron is searched for and updated accordingly. A couple of time-varying quantities drives the training algorithm: 1) a learning rate, $\eta(t)$, determines how much a sample affects vector positions; this quantity progressively decreases (typically in a linear function) when time increases. 2) A neighborhood function, $g(\rho,t)$ rewards the *winner* unit and its neighbours along the grid. The reward for the neighbouring units, too, decreases when time increases, hence the training strategy shifts from a spreading-activation pattern to a true winner-take-all schema. Figure 1 presents sample curves for the time-dependent quantities. The weight-update algorithm for SOM training can be outlined as follows:

Kohonen's Algorithm for Self-Organizing Maps

1. *For each* p-th input sample, $\mathbf{x}_p$; p =1,...,N_p
   - 1.a *Locate* the closest prototype (winner), $\mathbf{w}_{B(p)}$, such that:
     $$\left\|\mathbf{x}_p - \mathbf{w}_{B(p)}\right\| \le \left\|\mathbf{x}_p - \mathbf{w}_n\right\| \quad \forall n = 1,\ldots, N$$
   - 1.b *Update* the weights of the winner and those of its neighbours according to:
     $$\Delta\mathbf{w}_\rho = \eta(t)\cdot g(\rho,t)\cdot\left(\mathbf{x}_p - \mathbf{w}_{B(p)}\right)$$
     where $\mathbf{w}_\rho$ denotes a node laying at distance ρ in the neuron lattice.
   - 1.c *Increment* time
   - 1.d *Update* learning rate and neighborhood gain.

The principle of operation of Kohonen's model is that the number of prototypes is smaller than that of training samples; thus Vector Quantization involves a "compression" process. After training, each vector is assigned a specific portion of the sample set, and the overall vocabulary spans a tessellation of the data space.

At the same time, the final vector configuration is topologically consistent with the actual data distribution and facilitates a visual inspection of data coverage. In this sense, one can observe how training samples scatter among available prototypes (for example, by counting how many samples are covered by each vector).

4.2. Plastic Vector Quantization

The application of SOMs somehow implies that one estimates in advance the number of prototypes that best renders data distribution, let alone the problem of determining the suitable topology (the lattice of unit connections). In a huge, unexplored database, however, this may prove difficult to assess. Different and possibly more sophisticated techniques are required to investigate the intrinsic data dimensionality; the goal is to achieve a data representation that is both detailed enough to yield satisfactory accuracy and sufficiently simple to preserve some generalization ability. From a pattern-recognition perspective, the problem of selecting the best representation model reflects the so-called *bias-variance dilemma*.

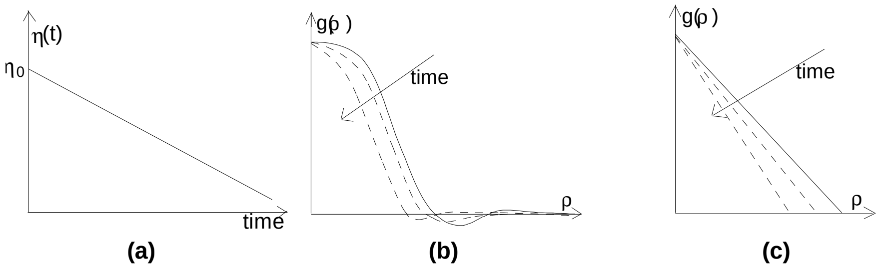


Fig. 1 – Variation in time for learning rate (a) and neighborhood width in the original Kohonen model (b) and in the usual version (c).

In unsupervised Vector Quantization, such a crucial issue is often tackled by *plastic* mechanisms [5, 6], enabling a network both to increase the number of neurons for improving data coverage (decreasing bias), and to remove neurons to enhance representation efficiency (decreasing variance). For the problem considered, we used a novel plastic model of VQ [7]. The algorithm is called "Plastic Neural Gas" because it exploits the Neural Gas algorithm [8] for vector positioning. Neural Gas follows a positioning strategy very close to Kohonen's algorithm, except that: 1) neurons are not topologically constrained, and 2) neighborhood is defined in the data space rather than in a fixed lattice. The neuron growing/pruning mechanism is driven by a local assessment of a neuron's placement. In practice, if a neuron's MSE exceeds a given threshold, a new unit is generated to improve coverage; conversely, if a unit fails to cover a data-space region (dead vector), it is pruned.

An important (and often disregarded) issue of plastic models concerns their generalization ability. In other words, uncontrolled plastic methods may tend to favor an accurate representation of training data to the detriment of generalization: the decrease of a cost (1) on training data does not necessarily involves a proportional improvement in the representation of the whole (unknown) sample domain. This basic problem is known as "overfitting", and, in the presented research, has been tackled by means of a cross-validation statistical technique. In each experiment, the data collection is split into a *training* set and a *test* set: the former is used to adjust a network's representation accuracy, whereas the latter provides an experimental assessment about a vocabulary's generalization effectiveness.

The dynamic Neural Gas model starts from a minimal network configuration (one neuron) and lets the network grow according to the plastic algorithm. Evaluating representation cost (1) on both training and test sets allows one to determine when a network has grown a sufficient number of prototypes: the algorithm stops when an improvement in training cost does not produce an enhancement of the test cost.

Although this stopping technique relies on empirical measurements, a number of thorough statistical theories provide analytical models for the generalization phenomenon. These models predict a network's test performance on the basis of training outcomes and sampling conditions; thus we checked the fitness of our experimental evidence to some well-known estimates in the literature, including Akaike's Information Criterion (AIC) [9], Rissanen's Minimum Description Length (MDL) criterion [10], and Vapnik's worst-case generalization theory for regression [11].

From a general perspective, the overall result of a Plastic Neural Gas training is equivalent to that of a SOM, as it yields a partitioning schema of represented data; the basic advantage lies in the method's adaptivity. On the other hand, plastic models may prove very expensive from a computational point of view, especially when applied to very complex databases.

5. Experimental results

The initial database validation included a subset of 2,000 samples out of the total number of 16,000 patients covered. The sample set was randomly chosen from the entire database; its limited size is mainly motivated by the preliminary stage of the research and by the huge computational cost involved by the validation process. From a general point of view, the main result is that there is no special cluster in the data. This is equivalent to assess that the sample is statistically well chosen.

5.1. Results from SOM-based analysis

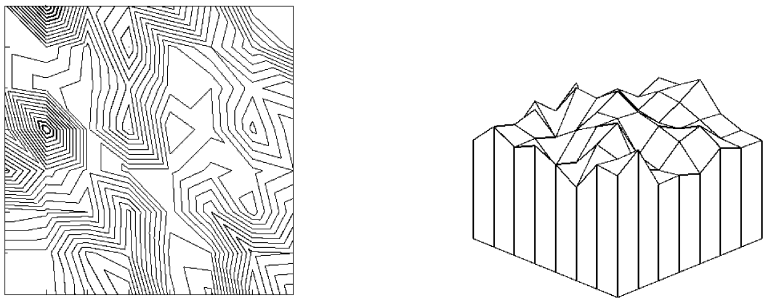
As said previously, the samples have been tested by means of 64 adaptive reference vectors. The average approximation error (average distance with respect to the range of the input values) is about 1.17%, that means a fairly good approximation even if we represented 2000 input patterns with a set of 64 reference vectors (number reduced to 64/2000 = 3.2 %). As no generalization performance was being evaluated in these preliminary tests, the entire sample set of 2,000 patients were used for the network training. Several test runs were performed to avoid bias from the algorithm's initial conditions; all runs, anyway, led to equivalent result.

Figure 2 shows graphically a typical result data density (i.e., the number of patients covered by each reference vector), and the proportion of data for each reference vector (right). The graph is a direct representation of the lattice square topology; neurons are arranged along the coordinate axis in the same order they appeared on the lattice grid. "Thick" areas indicate possible clusters, marked by neurons covering a relatively large number of samples as compared to other units.

The figures show a quasi-uniform distribution of samples among prototypes. In fact, further analysis will demonstrate in the next section that the number of prototypes (64) is smaller than the true data dimensionality; forcing the network to perform a coarser data compression may help evidence possible and well-marked natural groups. The absence of peaks in data coverage is a direct proof of the uniform patient distribution, and witnesses the correctness of the population sampling.

5.2. Results from Plastic Neural Gas

The training and test set for the experiments involving the plastic VQ model included 1,000 (randomly chosen) samples each. In all runs, the network was initialized with one unit and let grow as described in Section 4.2. A total number of 200 iterations of the Plastic Neural Gas algorithm were performed to assess the actual domain dimensionality. Figure 3 displays the progress of the training and test cost functions for two different experiments involving two different training/test set compositions. The arrow marks incipient



(a) (b)

Fig. 2 - Sample coverage results by a square 64-units Self-Organizing Map: (a) density representation; (b) Percent coverage.

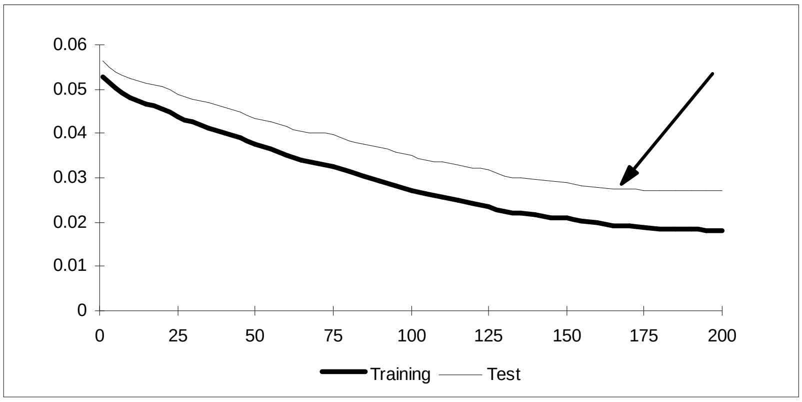


Fig. 3 - Sample curves for training and test cost versus number of units in plastic-model experiments.

overfitting indicated by a flattening test cost.

Experimental evidence from adaptive VQ indicates that the actual dimensionality of the considered data lies in the range [140,160]. Incidentally, this confirms that the previous test using SOMs have been performed in a data-compression situation.

The second research step with Plastic Neural Gas involved the characterization of the model's generalization performance according to statistical theories. Figure 4 presents a typical example of the ratio training cost/test cost, as experimentally observed (thick line) and as predicted by Akaike's AIC, Rissanen's MDL, and Vapnik's formula. Vapnik's model follows a worst-case approach and, as expected, predicts a much worse performance than experimentally measured. Rissanen's model still yields some underestimate, which is possibly due to the general conditions (although not universal, as in Vapnik's model) underlying their statistical analysis. Both graphs indicate a very good fit by

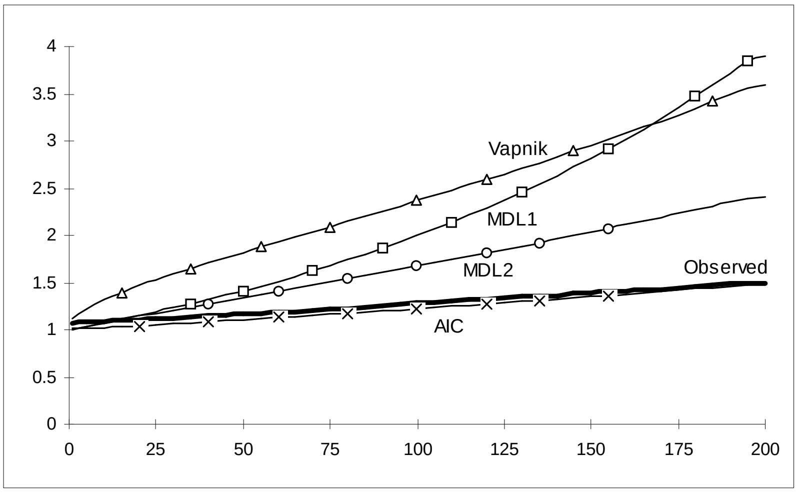


Fig. 4 – Fitness comparison of theoretical predictions: actual and predicted training/test ratios versus number of units.

Akaike's model. Remarkably, this is a confirmation of the sampling correctness, as Akakike's theory appears somehow more restrictive than the others, and implicitly assume a more accurate sampling of the domain under consideration.

References

[1] Hassager C, Christiansen G, "Assessment of bone mass," *IV International Symposium on Osteoporosis and Consensus Conference Proceedings*, 41-42, Hong Kong 1993.
[2] Johnell, "Fracture outcomes: consequences of osteoporosis for individuals and society," *IV International Symposium on Osteoporosis and Consensus Conference Proceedings*, 67-69, Hong Kong 1993.
[3] Martinetz T, Schulten K, "Topology representing networks," *Neural Networks*, 1994, vol.7, No.3, pp.507-522.
[4] Kohonen T, *Self organization and associative memories* (Springer Series in Information Sciences 8), Heidelberg:Springer, 1982.
[5] Fritzke B, "Let it grow – self-organizing feature maps with problem dependent cell structure," in Kohonen T et al. (Eds.) *Artificial neural networks*, 1991, North-Holland, pp. 403-408.
[6] Choi D-I, Park S-H "Self creating and organizing neural networks" *IEEE T Neural Netw*, 1994, vol.5, No.4, pp. 561-575.
[7] Ridella S, Rovetta S, Zunino R "Plastic Neural Gas for adaptive vector quantization," in preparation.
[8] Martinetz TM, Berkovich SG, Schulten KJ "'Neural Gas' network for vector quantization and its application to time-series prediction" *IEEE T Neural Netw*, 1993, vol.4, No.4, pp. 558-569.
[9] Akaike H, "A new look at the statistical model identification," *IEEE Trans. Autom.Control*, 1974, vol.AC-19, pp.716-723.
[10] Rissanen J "A universal prior for the integers and estimation by minimum description length," *Ann. Statist*., 1983, vol.11, pp.417-431.
[11] Vapnik V, *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag, New York, 1982.