LA STABILITÀ DELLE RETI NEURONALI NELL'OTTICA DI UNA STRATEGIA INTEGRATA DI ANALISI*

Cristina Davino e Domenico Vistocco

Dip.to di Matematica e Statistica – Università degli Studi di Napoli Federico II

Riassunto

Obiettivo del lavoro è l'introduzione di una strategia integrata di analisi per l'utilizzo delle reti neuronali in un contesto statistico. In particolare, viene affrontato il problema della stabilità sia dal punto di vista dei parametri che del campione. Relativamente al problema della stabilità dei pesi, viene proposto un approccio per analizzare quanto i pesi finali, anche se dipendenti dalla inizializzazione, si possono ritenere robusti. La stabilità del campione viene distinta in una stabilità interna legata alla presenza di osservazioni anomale, ed in una stabilità esterna, legata alla rappresentatività del campione. Le fasi relative alla selezione dell'architettura, alla definizione del modello e all'addestramento sono affrontate attraverso simulazioni su differenti modelli per poi procedere alla scelta di quello che presenta una buona performance nel rispetto del principio della parsimonia.

1. INTRODUZIONE

L'utilizzo delle reti neuronali sta avendo sempre più larga diffusione in ambiti tipici delle applicazioni statistiche. A fronte della iniziale diffidenza degli statistici verso tali tecniche è seguito un forte interesse verso di stesse. I primi contributi (Cheng e Titterington, 1994) (Ripley, 1994) sono stati rivolti ad esplorare le modalità di funzionamento delle reti neuronali, con particolare attenzione alla possibilità di replicare i tradizionali metodi statistici sfruttando questa differente

* *Il presente lavoro é stato finanziato dal Progetto Murst 2000 "Data Mining e analisi simbolica"; responsabile prof. N.C. Lauro.*

tecnologia. Questo tipo di approccio ha sì permesso di capire le forti potenzialità delle reti neuronali ma ha anche posto in risalto le limitazioni che caratterizzano tali modelli in applicazioni tipiche delle analisi statistiche.

Tali limiti hanno spinto gli statistici a collaborare con i ricercatori provenienti dai settori dell'informatica, in particolare l'intelligenza artificiale, al fine di sfruttare i vantaggi reciproci delle discipline, ovvero la potenza computazionale delle reti con l'attenzione alla trasparenza dei risultati e soprattutto all'interpretabilità degli stessi, caratteristiche queste peculiari della metodologia statistica.

Alcuni contributi hanno cercato di inquadrare le reti neuronali in una strategia integrata di analisi (Davino *et al.*, 1997) (Lauro *et al.*, 1999) in maniera da ridurre al minimo la discrezionalità del ricercatore nelle varie fasi e di porre l'attenzione a particolari momenti dell'analisi con l'obiettivo di limitare le debolezze di questo tipo di modelli.

In questo lavoro si riprende l'approccio integrato alle reti neuronali dedicando particolare attenzione alla parte relativa alla stabilità sia dal punto di vista dei parametri del modello che rispetto al campione di dati utilizzato per la definizione dello stesso. Relativamente al problema della stabilità dei pesi, viene proposto un approccio per analizzare quanto i pesi finali, anche se dipendenti dalla inizializzazione, si possono ritenere robusti. La stabilità del campione viene distinta in stabilità interna, relativa alla presenza di osservazioni anomale nel campione, e in stabilità esterna, relativa alla rappresentatività del campione stesso. Un'analisi più approfondita è stata effettuata anche per le altre fasi dell'analisi, in particolare per le fasi di selezione dell'architettura, definizione del modello ed apprendimento sono state effettuate varie simulazioni su differenti modelli per poi procedere alla scelta di quello che presenta una buona performance tenendo conto contemporaneamente del principio della parsimonia.

Il lavoro è strutturato nelle seguenti sezioni:

- nel paragrafo successivo si illustra l'evoluzione nell'atteggiamento dei ricercatori statistici nei confronti delle reti neuronali, mettendo in risalto i punti in comune alle due discipline e le divergenze tra le stesse;
- nel terzo paragrafo si illustra la strategia di analisi proposta da Davino et al. (1997). L'analisi dei singoli passi della strategia si serve di un'applicazione di riferimento su dati reali allo scopo di illustrare praticamente la modalità di analisi suggerita; il passo della strategia riguardante la stabilità è quello cui si dedica maggiore attenzione;
- alcune considerazioni conclusive sono riportate nel quarto paragrafo.

2. RETI NEURONALI E I METODI STATISTICI: *PRO E CONTRO*

Le reti neuronali sono state sviluppate nell'ambito delle scienze cognitive per studiare ed approfondire l'attività delle reti neuronali biologiche e per imitarne le enormi potenzialità. Le reti neuronali sono nate quindi come strumento di calcolo caratterizzato da un'elaborazione parallela e non sequenziale dell'informazione e dalla flessibilità e capacità di apprendere dall'esperienza, attività questa tipica del cervello umano. In breve tempo le reti neuronali hanno trovato applicazione in molti campi ritenuti non praticabili dalle metodologie tradizionali in quanto caratterizzati dalla presenza di molti dati, spesso incompleti o inquinati da errore e non registrati su archivi ma continuamente aggiornati. In tale contesto applicativo le reti neuronali sono state utilizzate con successo in svariati settori: il settore delle telecomunicazioni, ad esempio, per l'analisi dei segnali e per l'analisi delle interferenze; in campo militare, per il riconoscimento degli obiettivi, per la navigazione automatica e per il riconoscimento dei segnali radar; in campo industriale, per l'automazione dei robot e per la creazione di sistemi di controllo; in campo ambientale, per le previsioni meteorologiche.

Allo sviluppo verticale che nell'arco di alcuni decenni ha caratterizzato l'evoluzione delle reti neuronali dal semplice neurone artificiale di McCulloch e Pitts (1943) alle reti non supervisionate (Kohonen, 1982) fino alle reti supervisionate multistrato (Rumelhart *et al.*, 1986), è seguito, dagli inizi degli anni '80, un ben più rapido e non sempre giustificato sviluppo orizzontale che ha avuto come principale conseguenza l'uso sempre più frequente delle reti neuronali nei più diversi campi di applicazione (medicina, finanza, credit scoring, ambiente, assicurazioni) e quindi l'inevitabile sovrapposizione con i metodi statistici che in quei campi erano tradizionalmente utilizzati.

L'atteggiamento iniziale degli statistici riguardo i modelli a reti neuronali è stato in principio ostile: diverse sono state le polemiche, anche in letteratura, sulla effettiva originalità e necessità di utilizzare le reti neuronali piuttosto che i tradizionali metodi in un ambito proprio delle applicazioni statistiche, forse lontano da quello originario delle reti e sicuramente più rigoroso nella definizione di ipotesi e nella richiesta di trasparenza riguardo i processi che portano al conseguimento dei risultati finali.

A tale proposito di seguito riportiamo alcune considerazioni del mondo scientifico.

Hecht-Nielsen (1990) afferma che il "principale vantaggio" della disciplina che si occupa di reti neuronali (*neurocomputing*) rispetto alla statistica risiede nella possibilità che tale disciplina ha di attingere da "diverse fonti di ispirazione". La

mancanza di un apporto metodologico innovativo da parte delle reti neuronali induce quindi l'autore a ritenere che, dopo l'apporto di Gauss alla statistica, ci siano stati solo modesti progressi: *"The main advantage of neurocomputing over statistics seems to be neurocomputing's ability to draw from many more sources of inspiration (...) for its methods than are exploited by workers in statistics. In essence, in terms of its everyday practice, there has only been modest progress in regression analysis since the days of Gauss"*.

Per alcuni autori (E. Robert Tisdale, *user group* comp.ai.neural-nets, 6 novembre 1994) le reti neuronali sembra abbiano apportato più che contributi innovativi solo confusione e una diversa terminologia: *"There is nothing new or magical about neural networks. Artificial neural networks are sometimes used as "black box models" with lots of parameters which can be adjusted to approximate the response of any system provided that there are sufficient data to estimate the parameters accurately. The method used to adjust these parameters have rarely changed at all since they were invented by Karl Friederich Gauss about 200 years ago. Neural network research seems to have contributed almost nothing at all save for jargon and confusion"*.

In particolare negli ultimi anni molte tecniche statistiche sono state riprodotte utilizzando i modelli neuronali per cui l'unica differenza tra i due approcci è sembrata essere la terminologia.

Sarle (1996) fornisce una traduzione accurata della terminologia neuronale in quella statistica (Figura 1) mostrando come la letteratura statistica e quella relativa alle reti neuronali in alcuni casi esprimano gli stessi concetti con una terminologia diversa e, in altri casi, gli stessi termini vengano utilizzati con significati diversi.

Statistica		**Reti Neuronali**
Modello	⟺	Architettura
Osservazioni	⟺	Pattern
Variabile indipendente	⟺	Input
Variabile dipendente	⟺	Output
Stima	⟺	Apprendimento
Parametri	⟺	Pesi
Regressione, Seq. Binaria, An. Discriminante	⟺	Reti supervisionate
Cluster Analysis	⟺	Reti no supervisionate

Fig. 1: Un confronto di terminologia.

L'utilizzo delle reti neuronali in contesti tipici dei metodi statistici, inoltre, induce a tenere presente oltre ai problemi tipici delle reti, quali la selezione dell'architettura e l'apprendimento lento su macchine sequenziali, alcune peculiarità delle stesse quali la non univocità della soluzione finale e la mancanza di trasparenza nell'interpretazione dei risultati, che costituiscono problemi determinanti in statistica.

Il confronto tra metodi statistici e reti neuronali (Bishop, 1995), inizialmente applicativo e successivamente metodologico, ha consentito di dissipare, almeno in parte, l'iniziale diffidenza del mondo statistico verso tali modelli riconoscendo, così, alle reti neuronali il merito di riuscire a risolvere problemi complessi anche in assenza di ipotesi distribuzionali.

Anzichè effettuare solamente un semplice confronto fra reti neuronali e modelli statistici compaiono in letteratura i primi contributi in cui viene posto l'accento sulle possibili sinergie che possono derivare da un utilizzo congiunto delle due metodologie: *" ... a novel integrated approach emerges which stresses both flexibility (contribution of neural nets) and interpretability (contribution of tatistical modeling)"* (Ciampi e Lechevallier, 1995).

Per individuare il posizionamento dei modelli neuronali nel quadro più generale di tutti i metodi statistici si può fare ricorso ad una rappresentazione grafica in cui i diversi metodi si posizionano in uno spazio a tre dimensioni. La prima dimensione indica la **complessità** del fenomeno oggetto di studio: i metodi che analizzano una sola variabile (univariati) si oppongono a quelli che analizzano più variabili (multivariati). La seconda dimensione si riferisce al **grado di incertezza** del fenomeno, per cui i modelli deterministici si contrappongono ai modelli probabilistici o stocastici. La terza dimensione è definita dall'eventuale **carattere temporale** del fenomeno che implica il ricorso a metodi dinamici, se tale caratteristica è presente, o statici, in caso di una sua assenza.

I vertici del cubo in Figura 2 rappresentano otto possibili combinazioni di metodologie con diverse caratteristiche. Il *vertice 1* rappresenta la situazione più semplice: una tecnica deterministica, univariata, statica (per esempio un indice di statistica descrittiva). I vertici 2, 3 e 5 pur riferendosi a fenomeni univariati sono caratterizzati da diverse combinazioni del fattore complessità e di quello incertezza. In particolare il *vertice 2* rappresenta una metodologia deterministica, univariata ma dinamica (per esempio un metodo di analisi delle serie storiche univariate descrittive o classiche). Il *vertice 3* rappresenta una metodologia probabilistica, univariata e statica (per esempio l'inferenza statistica) mentre il *vertice 5* una metodologia probabilistica, univariata, ma dinamica (per esempio i modelli di analisi delle serie storiche secondo l'approccio stocastico). Gli altri quattro vertici,

situati nella parte alta del cubo, si riferiscono a fenomeni complessi. In particolare il *vertice 4* (tecnica multivariata, statica e deterministica) rappresenta sicuramente la famiglia dei metodi di analisi multidimensionale dei dati, secondo l'approccio della scuola francese. L'aggiunta del fattore tempo *vertice 7* determina il ricorso ai metodi di analisi multidimensionale a più vie. Il *vertice 6* si riferisce invece a modelli probabilistici complessi ma statici (per esempio i metodi di analisi multivariata dei dati secondo l'approccio della scuola inglese). La classe di problemi più difficili da risolvere è rappresentata sicuramente dal *vertice 8* che rappresenta i problemi caratterizzati da molte variabili, molta incertezza e dinamicità. I modelli neuronali si posizionano nella parte alta del cubo (vertici 4, 6, 7, 8) rappresentato in Figura 2 insieme ad altre tecniche tipiche dell'Intelligenza Artificiale (algoritmi genetici (Goldberg, 1989) (Rizzi, 1998), sistemi esperti (Buchanon e Shortliffe, 1984), (Stevens, 1984), logica fuzzy (Zadeh, 1977), ecc.). In particolare le reti neuronali cercano di colmare la lacuna della mancanza di modelli in tale area potendosi quindi considerare non come uno strumento sostitutivo dei modelli già esistenti ma come uno strumento di integrazione in quelle situazioni in cui i metodi tradizionali non riescono a fornire risultati soddisfacenti.

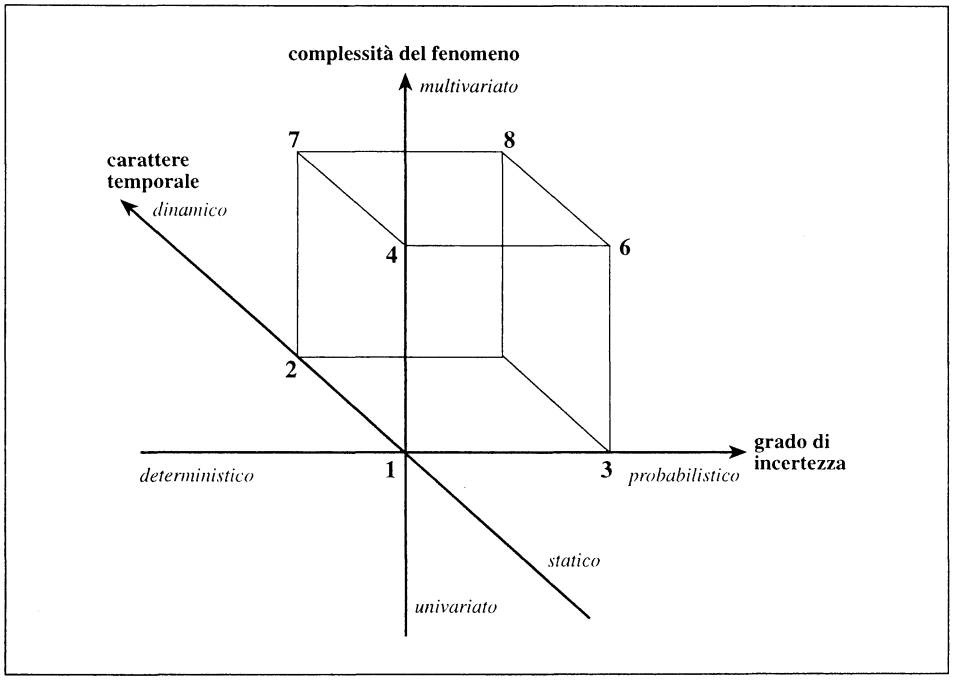

Fig. 2: Rappresentazione dei metodi statistici.

3. UN APPROCCIO STATISTICO ALLE RETI NEURONALI

L'utilizzo delle reti risulta particolarmente significativo ed utile quando inquadrato in un'ottica di interazione con la statistica al fine di sfruttarne le possibili sinergie (Davino *et al.*, 1997). Le reti neuronali possono quindi essere utilizzate come un momento di una più generale strategia di analisi che potremmo definire "integrata" (Davino *et al.*, 1997) (Lauro *et al.*, 1999) e che preveda, accanto ad una fase di *pre-processing* delle variabili, l'utilizzo di metodi statistici per la selezione dell'architettura, per la verifica della capacità di generalizzazione della rete e della stabilità dei risultati ottenuti (Figura 3).

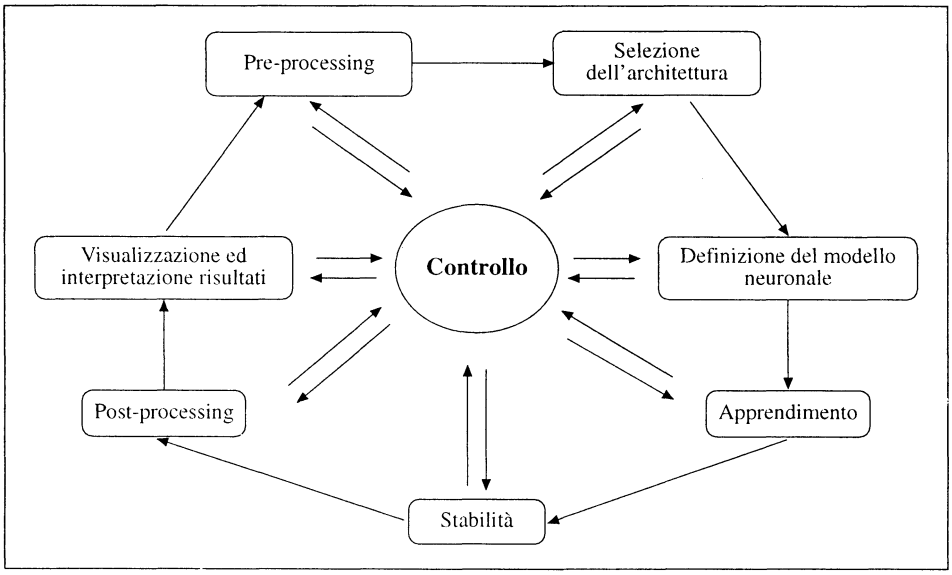

Fig. 3: Una strategia di utilizzo statistico delle reti neuronali.

La prima fase di tale strategia di analisi è il **pre-processing** dei dati, vale a dire un'accurata analisi descrittiva degli stessi, la loro trasformazione e codifica secondo le esigenze dei modelli neuronali ed una eventuale riduzione delle variabili attraverso una tecnica statistica quale un'Analisi Fattoriale o un Modello Grafico. Una fase di pre-training può risultare utile per una scelta più appropriata dei pesi iniziali. La **definizione dell'architettura** della rete è un momento cruciale della strategia di analisi in quanto da essa dipende la capacità di generalizzazione di una rete. Il problema della scelta del numero di strati intermedi e del numero di neuroni in ciascuno di essi, in genere affrontato provando diverse architetture e scegliendo tra queste quella che porta ad un tasso di errore più basso, trova un valido supporto,

oltre che nei metodi di segmentazione binaria, anche nei cosiddetti metodi di pruning e negli algoritmi con apprendimento adattivo. Dopo la **scelta del modello** e l'**addestramento** della rete si rende necessaria una valutazione della performance della rete ed una verifica della **stabilità dei risultati** ottenuti. L'interpretabilità di questi ultimi richiede una fase finale di **post-processing** dei dati.

Le varie fasi della strategia vengono descritte nei paragrafi seguenti tramite un'applicazione su dati reali, mettendo in risalto l'importanza che assume in tale contesto l'approccio statistico. In particolare, maggiore attenzione verrà dedicata alla fase della stabilità dei risultati che risulta influenzata sia dall'inizializzazione dei pesi che dalla selezione dell'architettura che dal campione utilizzato per l'apprendimento.

3.1 I DATI

L'applicazione è stata condotta utilizzando un estratto del data set "*German credit data*" fornito dal prof. Hofmann e disponibile sul sito ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/. Si tratta di un data set tipico delle applicazioni di credit scoring: su un campione di 1000 clienti cui una banca ha concesso un prestito vengono rilevate informazioni anagrafiche e bancarie insieme all'esito finale del prestito (restituito o evaso). L'obiettivo dell'applicazione è costruire un modello neuronale utilizzabile per prevedere con il più alto grado di attendibilità, per un nuovo cliente che faccia richiesta di prestito, l'esito dello stesso. L'estratto del "*German credit dataset*" preso in considerazione si riferisce a 1000 clienti sui quali sono state osservate le seguenti variabili (in parentesi, per ciascuna variabile, sono riportate le categorie):

1. importo c/c (<0 marchi tedeschi; 0-200 marchi tedeschi; >=200 marchi tedeschi; no conto corrente)
2. situazione crediti (no crediti; crediti passati pagati; crediti attuali pagati; ritardo pagamento; altri crediti)
3. impiego (disoccupato; operaio non specializzato; operaio specializzato; dirigente)
4. telefono (sì; no)
5. durata del prestito in mesi (<1 anno; 1-2 anni; +2anni)
6. importo credito (basso; medio; alto)
7. % indebitamento sul reddito (1-2%; 3-4%)
8. età (19-28 anni; 29-38 anni; >39 anni)
9. motivo (auto nuova; auto usata; mobili; elettrodomestici; affari; altro motivo)
10. libretto risparmio (<100 marchi tedeschi; 100-500 marchi tedeschi; >500 marchi tedeschi; no libretto di risparmio)

11. genere (maschi; femmine)
12. numero di rapporti attivi con la banca (1 rapporto attivo; +1 rapporto attivo)
13. esito del prestito (restituito; insoluto)

3.2 IL PRE-PROCESSING

Il primo aspetto della fase di *pre-processing* che occorre prendere in considerazione riguarda la definizione del ruolo giocato dalle variabili analizzate all'interno della rete: le 12 variabili esplicative verranno rappresentate nello strato di input e la variabile di risposta nello strato di output. Una volta definita la rappresentazione delle variabili occorre riflettere sulla loro natura: le variabili qualitative e quelle quantitative, indipendentemente dal ruolo all'interno della rete, vengono infatti rappresentate con un numero diverso di neuroni. Le variabili quantitative, siano esse esplicative o dipendenti, vengono rappresentate con un neurone. L'applicazione presentata prende in considerazione tutte variabili qualitative che, siano esse esplicative o dipendenti, vengono rappresentate in maniera binaria utilizzando un numero di neuroni in input (output) per ciascuna variabile esplicativa (di risposta) pari al numero di modalità della variabile stessa. La matrice dei dati in codifica ridotta di dimensione 1000 × 13 viene trasformata nella matrice in codifica disgiuntiva completa di dimensioni 100 × 42. In realtà non è necessario che il numero di neuroni utilizzati per ciascuna variabile sia esattamente uguale al numero delle sue modalità: è possibile infatti eliminare una modalità, e quindi un neurone, per ciascuna variabile qualitativa in quanto tale modalità risulta univocamente determinata dalle restanti. La topologia della rete sarà quindi costituita da uno strato di input con *40-12=28* neuroni ed uno strato di output con *2-1=1* neurone. Si è ottenuto, così, un risparmio pari a 42-29=13 neuroni.

Una volta definita la rappresentazione delle variabili all'interno della rete è spesso necessario procedere ad alcune trasformazioni delle stesse: la standardizzazione delle variabili indipendenti viene in genere effettuata in caso di variabili espresse in diverse unità di misura, la standardizzazione delle variabili dipendenti consente di ridurre la scala dei pesi iniziali e garantisce una migliore performance; la normalizzazione (Smith, 1993), sia delle variabili indipendenti che di quelle dipendenti, può risultare conveniente quando si scelgono alcune funzioni di trasferimento, in particolare per quelle sigmoidali come la logistica[1].

---

[1] *Le funzioni di tipo sigmoidale hanno infatti andamento asintotico per i valori al di fuori di un dato intervallo centrale: la normalizzazione consente di ridurre la perdita di informazione per i valori di input al di fuori del suddetto intervallo.*

Le variabili indipendenti sono state normalizzate nel range [-1; 1] e la variabile di risposta nel range [0.1; 0.9] secondo la seguente formula:

$$ValNorm = ValMin + \left(\frac{Val - Min}{Max - Min} (ValMax - ValMin) \right)$$

dove *ValMin* è il limite inferiore del range di normalizzazione, *ValMax* è il limite superiore del range di normalizzazione, *Min* e *Max* rispettivamente il minimo ed il massimo della variabile da normalizzare e *Val* il valore della variabile da normalizzare.

3.3 LA SELEZIONE DELL'ARCHITETTURA

La scelta di un'appropriata complessità della rete influenza in maniera determinante la capacità di generalizzazione della stessa, ovvero la capacità di trattare dati che, pur appartenendo allo stesso spazio campionario dei dati su cui la rete è stata addestrata, presentano, rispetto a questi, delle differenti caratteristiche.

Tra le diverse strategie proposte in letteratura per la definizione dell'architettura di una rete si è pensato di addestrare più reti con diverse architetture e scegliere quella che consente di ottenere la migliore combinazione di performance sul training set (comprendente il 70% dei casi) e sul test set (comprendente il 30% dei casi), tenendo conto del principio della parsimonia per discriminare tra situazioni simili. A tale proposito sono state addestrate 135 reti con uno strato nascosto, ottenute variando il numero di neuroni in tale strato (da 1 a 5), le funzioni di trasferimento (lineare, logistica, tangente) e l'algoritmo di apprendimento (backpropagation, backpropagation con momentum, backpropagation con tasso di apprendimento adattivo, backpropagation con tasso di apprendimento adattivo e momentum, algoritmo Fletcher-Reeves di gradiente coniugato, algoritmo Polak-Ribiére di gradiente coniugato, algoritmo Powell-Beale di gradiente coniugato, algoritmo di gradiente coniugato scalato, algoritmo quasi Newton, algoritmo quasi Newton di Broyden, Fletcher, Goldfarb e Shanno, algoritmo Levenberg-Marquardt, algoritmo Levenberg-Marquardt con parametro per la riduzione di memoria) (Hagan *et al.*, 1996).

In Figura 4 sono riportate, in base alla percentuale di corretta classificazione sul training set e sul test set, le architetture di rete tra cui è stata effettuata la selezione. La scelta dell'architettura da utilizzare ha seguito il duplice criterio di massimizzare la capacità di generalizzazione della rete (combinazione ottimale della corretta classificazione sul training e sul test set) minimizzando i costi

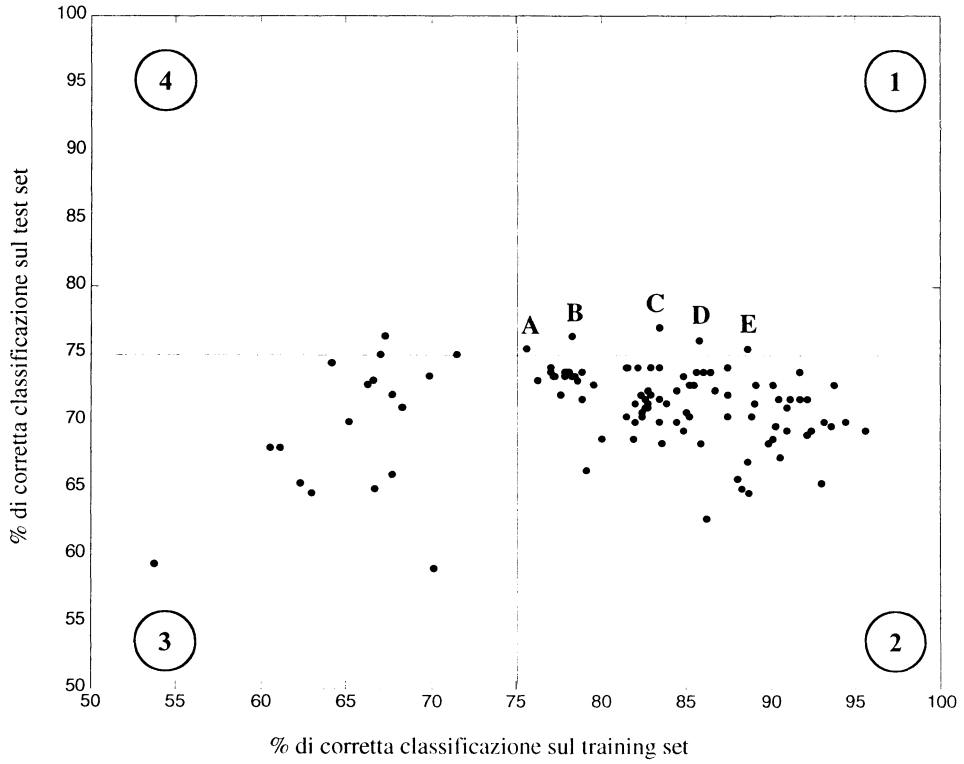

Fig. 4: Selezione dell'architettura.

computazionali (principio della parsimonia). Come soglia di divisione tra i quadranti in si è utilizzata una percentuale di corretta classificazione pari al 75% su entrambi i campioni a disposizione. Il quadrante 1 rappresenta la zona più interessante per la selezione del modello finale, in quanto contiene le reti con una percentuale di corretta classificazione superiore al 75% sia sul training set che sul test set. Gli altri quadranti riportano percentuali più alte solo sul training set (quadrante 2), solo sul test set (quadrante 4) o basse su entrambi (quadrante 3). Le caratteristiche dei modelli (algoritmo di apprendimento, funzione di trasferimento e nr. di neuroni nello strato intermedio, nr. totale di pesi, percentuale di corretta classificazione sul training set e sul test set) che ricadono nel quadrante 1 sono elencate nella Tabella 1.

Tab. 1: Caratteristiche dei modelli migliori.

Rete	Algoritmo di apprendimento	Funzioni di trasferimento	n.ro neuroni strato intermedio	n.ro totale di pesi	% di corretta classificazione (training set)	% di corretta classificazione (test set)
A	Backpropagation con tasso di apprendimento adattivo	Lineare	4	118	76%	75%
B	Algoritmo Powell-Beale di gradiente coniugato	Tangente	1	31	78%	76%
C	**Algoritmo Fletcher-Reeves di gradiente coniugato**	**Tangente**	**2**	**60**	**83%**	**77%**
D	Algoritmo Fletcher-Reeves di gradiente coniugato	Tangente	4	118	86%	76%
E	Algoritmo Powell-Beale di gradiente coniugato	Tangente	3	89	89%	75%

La selezione finale del modello, seguendo i principi sopra riportati, ha condotto alla scelta della rete **C** con 28 unità di input (determinato dal numero di variabili esplicative), con uno strato nascosto con due unità e con un neurone nello strato di output (determinato anch'esso a priori dalla presenza di un'unica variabile dipendente).

3.4 LA STABILITÀ

Lo studio della stabilità dei risultati di un modello rappresenta una fase cruciale dal punto di vista statistico ma non altrettanto cruciale per i modelli a rete neuronali. In letteratura, infatti, i contributi relativi a questo aspetto sono abbastanza recenti (Tibshirani, 1996). Si tratta spesso di contributi rivolti a particolari momenti di questa fase sicuramente non paragonabili all'attenzione che la letteratura statistica ha rivolto in passato all'argomento. L'interesse degli statistici verso le reti neuronali ha però posto in risalto il problema anche nella letteratura specialistica sull'argomento (Moody, 1994).

L'analisi della stabilità può essere divisa in due momenti, a seconda del punto di interesse cui ci si rivolge: **stabilità dei pesi** e **stabilità del campione**. Entrambe le fasi studiano la dipendenza dei risultati del modello dai parametri caratteristici dello stesso. La prima fase è fondamentalmente rivolta allo studio della dipendenza del modello dai pesi e in particolar modo dalla discrezionalità associata agli stessi: si tratta di uno studio di sensitività dei risultati rispetto alle possibili scelte a disposizione in fase di inizializzazione dei pesi, che possono determinare il valore finale dei pesi stessi. Lo studio della stabilità del campione, verso cui già in passato si riscontra maggiore attenzione in letteratura nei contributi rivolti allo studio del

fenomeno dell'overfitting (Sarle, 1995), riguarda invece la dipendenza dei risultati dalla scelta riguardante il campione di dati disponibili. Una trattazione dettagliata degli aspetti relativi a ciascuna fase, sia a livello metodologico che rispetto alla particolare applicazione, è riportata nei due sottoparagrafi seguenti.

3.4.1 LA STABILITÀ DEI PESI

Lo studio della stabilità dei pesi è rivolto all'analisi della dipendenza dei risultati del modello dalle scelte che il ricercatore deve compiere in fase di taratura dei parametri dello stesso. Nel considerare questo aspetto l'obiettivo è stabilire quanto i pesi finali ottenuti, seppur dipendenti dalla scelta iniziale, si possono ritenere robusti (cambiando le scelte in quale misura i parametri sono differenti e soprattutto i risultati finali del modello variano).

Selezionata la rete **C** in tabella 1 secondo i criteri esposti nel paragrafo precedente, si è proceduto a valutare la stabilità dei pesi finali eseguendo diverse inizializzazioni e confrontando la percentuale di corretta classificazione ottenuta sul training set per la rete **C** con le percentuali ottenute in seguito alle simulazioni. Nel caso in esame sono state effettuate mille inizializzazioni usando l'algoritmo di inizializzazione di Nguyen-Widrow (Hagan *et al.,* 1996). Tale scelta è stata dettata dalla pratica comune nelle applicazioni sulle reti neuronali: altre possibili modifiche potrebbero riguardare l'uso di differenti funzioni per la generazione dei valori iniziali da assegnare ai pesi. Una volta terminato l'addestramento si valuta la percentuale di corretta classificazione sul training set, in maniera da ottenere una distribuzione empirica di tali percentuali. Il valore della percentuale di corretta classificazione ottenuto considerando la rete selezionata viene confrontato con l'intervallo di confidenza della distribuzione empirica centrato sulla media e di ampiezza pari ad uno scarto quadratico medio e comprendente circa il 68% della distribuzione dei valori. In Figura 5 è riportato l'andamento della distribuzione della percentuale di corretta classificazione per le differenti inizializzazioni e gli estremi ([78,6%, 85,3%]) dell'intervallo di cui sopra. La percentuale di corretta classificazione ottenuta sulla rete scelta è pari all'83% (Tabella 1), valore che ricade all'interno dell'intervallo empirico calcolato, comprovando quindi la stabilità dei risultati ottenuti relativamente ai pesi.

Il secondo punto di analisi, stabilità dei parametri finali della rete rispetto alle scelte prese dal ricercatore in fase di taratura del modello, è rivolto a valutare l'effettiva robustezza rispetto alle condizioni iniziali fissate per l'applicazione di algoritmi. A tale proposito si è indagato su un'eventuale relazione esistente tra le inizializzazioni e le performance del modello ad esse associate, attraverso un metodo di discriminazione non parametrica. Il metodo scelto è la segmentazione binaria, usando come variabili esplicative i valori dei pesi nelle 1000 inizializzazioni

effettuate e come variabile di risposta le percentuali di corretta classificazione ad esse associate. La variabile di risposta è stata discretizzata utilizzando tre classi corrispondenti alle tre diverse regioni definite dall'intervallo di confidenza in Figura 5:

- Classe 1, percentuali di corretta classificazione minori di 78,6%;
- Classe 2, percentuali di corretta classificazione comprese tra 78,6% e 85,3%;
- Classe 3, percentuali di corretta classificazione maggiori di 85,3%.

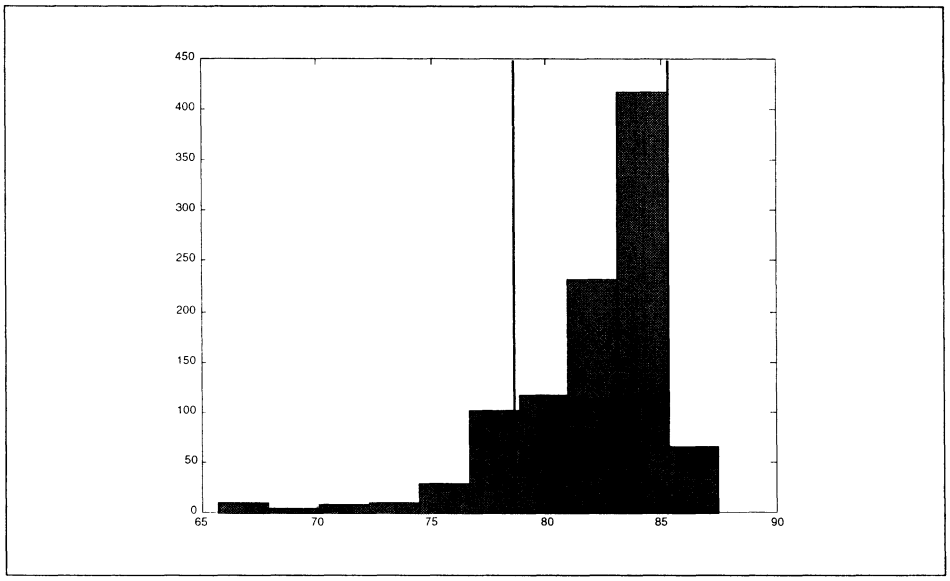

Fig. 5: Distribuzione della % di corretta classificazione per 1000 inizializzazioni casuali dei pesi.

La seconda classe è ovviamente la classe di interesse anche se va tenuto presente che la terza classe, pur presentando valori al di fuori dell'intervallo di confidenza empirico costruito in precedenza, racchiude modelli con migliori capacità predittive.

La segmentazione binaria, al pari di altri metodi di discriminazione, determina la regola di classificazione su una parte del campione (campione di apprendimento) utilizzando la restante parte (campione di test) per valutarne la capacità previsiva. In Figura 6 è riportato l'albero binario ottenuto dall'applicazione di uno dei metodi più diffusi di segmentazione binaria: il CART (Breiman *et al.*, 1984) sul campione di apprendimento costituito da 700 inizializzazioni. La radice dell'albero rappresenta il campione iniziale, i nodi intermedi sottoinsiemi di tale campione

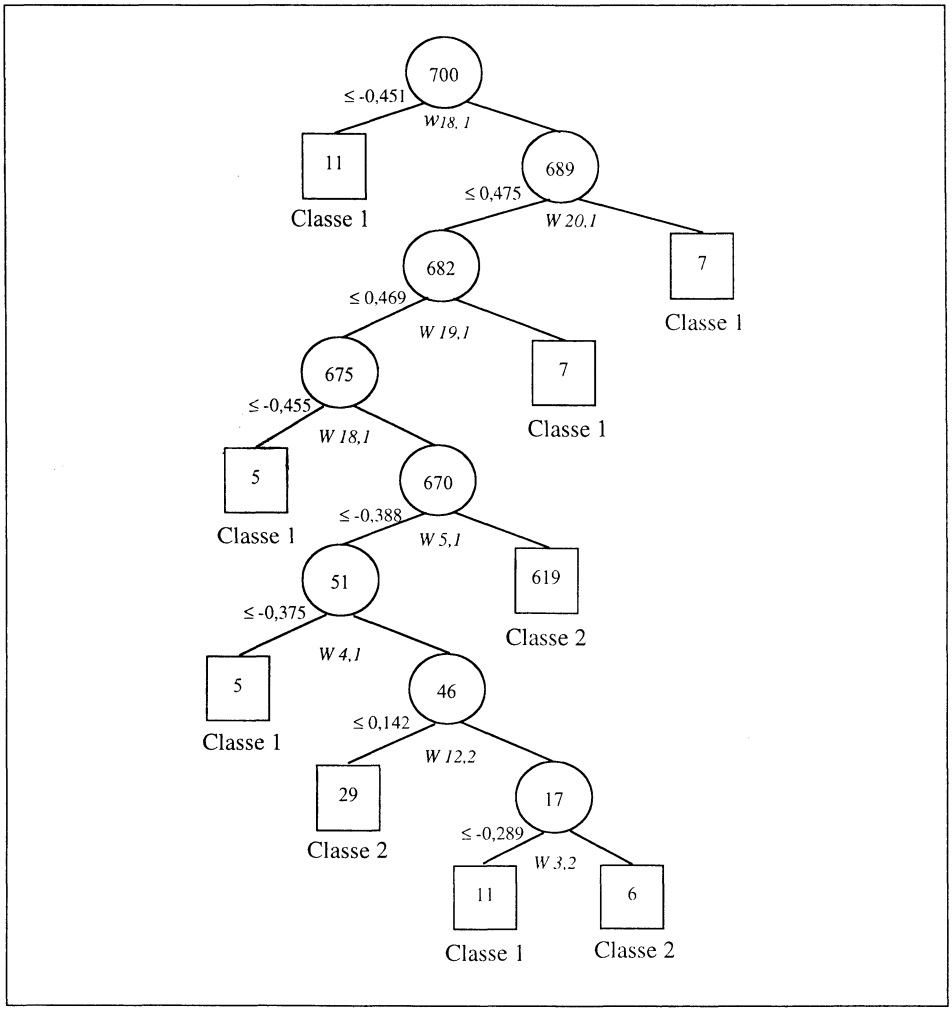

Fig. 6: L'albero binario.

ottenuti sulla base dei valori delle variabili esplicative e i nodi terminali le "foglie dell'albero" cioè nodi non divisibili che identificano individui appartenenti ad una delle tre classi precedentemente definite. A tale proposito, dalla figura 6 risulta che tutti i casi assegnati a priori alla classe 3, sono stati mal classificati a posteriori. In realtà, il 92% di tali casi è stato attribuito alla classe 2, evitando così sovrastime della percentuale di corretta classificazione. La regola di classificazione creata ha fornito una percentuale di corretta classificazione sul campione di apprendimento pari all'81% e sul campione test pari al 72%. Tali risultati possono considerarsi

soddisfacenti e permettono di ritenere che esiste, come previsto, una relazione tra pesi iniziali e performance della rete.

La regola di classificazione può essere utilizzata anche a fini previsionali su casi differenti da quelli utilizzati per la sua creazione e di cui non è nota la classe di appartenenza. Tale modalità di utilizzo avviene facendo "scivolare" i casi in questione all'interno dell'albero e percorrendone i rami appropriati sulla base dei valori delle variabili associate alle divisioni fino a giungere ad un nodo terminale che determina la classe prevista. Il caso da classificare è costituito dal vettore dei pesi di partenza della rete **C**: pur essendo nota la percentuale di corretta classificazione associata a tali pesi (83%), questo permette di valutare la stabilità dei pesi di partenza. Tale vettore viene attribuito alla classe 2 (percentuali di corretta classificazione comprese tra 78,6% e 85,3%) consentendo di riscontrare una coerenza con la reale percentuale di corretta classificazione (83%).

L'applicazione di un tale modello di discriminazione può essere rivolto all'approfondimento dell'influenza che i singoli pesi esercitano sulla performance finale considerando, in questo caso, i pesi ottenuti dopo l'addestramento partendo da differenti inizializzazioni: i pesi con maggiore potere discriminante sono quelli che l'algoritmo di segmentazione seleziona come variabili di divisione dei nodi intermedi. In quest'ottica, ulteriori sviluppi dell'approccio proposto potrebbero essere rivolti a ridurre il tipico effetto a scatola nera che ancora affligge i modelli a reti neuronali.

3.4.2 LA STABILITÀ DEL CAMPIONE

Lo studio della stabilità del campione è rivolto a valutare quanto la performance del modello dipende dal particolare campione utilizzato per la messa a punto dello stesso. A tale scopo il primo approccio al problema è stato il ricorso alla tecnica nota in letteratura come *stopped training* (Smith, 1993). Si tratta di una procedura che procede alla divisione del dataset a disposizione in due od eventualmente tre sottoinsiemi, vale a dire training set o campione di addestramento, validation set o campione di validazione e test set o campione di test. Il primo campione viene utilizzato per l'addestramento della rete: l'algoritmo di apprendimento mirerà quindi a minimizzare l'errore su tale campione. Il secondo campione, validation set, viene utilizzato per cercare di ridurre il fenomeno dell'overfitting mentre il terzo (quando la numerosità iniziale permette di ottenerlo) è utilizzato per testare la performance del modello su un dataset diverso prima del suo effettivo utilizzo. La procedura di stopped training è molto semplice ed intuitiva: la rete viene addestrata sul campione di apprendimento per un determinato numero di epoche; successivamente i pesi vengono congelati e si verifica la capacità predittiva/

classificatoria del modello sul validation set per poi ripartire nuovamente con l'addestramento. L'osservazione congiunta dell'andamento dell'errore sui due dataset permette di bloccare l'addestramento quando l'errore sul validation set, che non influenza la modifica dei pesi ma viene semplicemente usato per testare il modello ottenuto fino a quel determinato punto, inizia a salire: ulteriori epoche di addestramento potrebbero portare ad un eccessivo adattamento del modello al campione di addestramento pena una scarsa capacità predittiva/classificatoria su un campione differente. Un andamento tipico dell'errore sul training set e sul validation set è riportato in Figura 7, dove si pone in evidenza un potenziale punto dove arrestare la procedura.

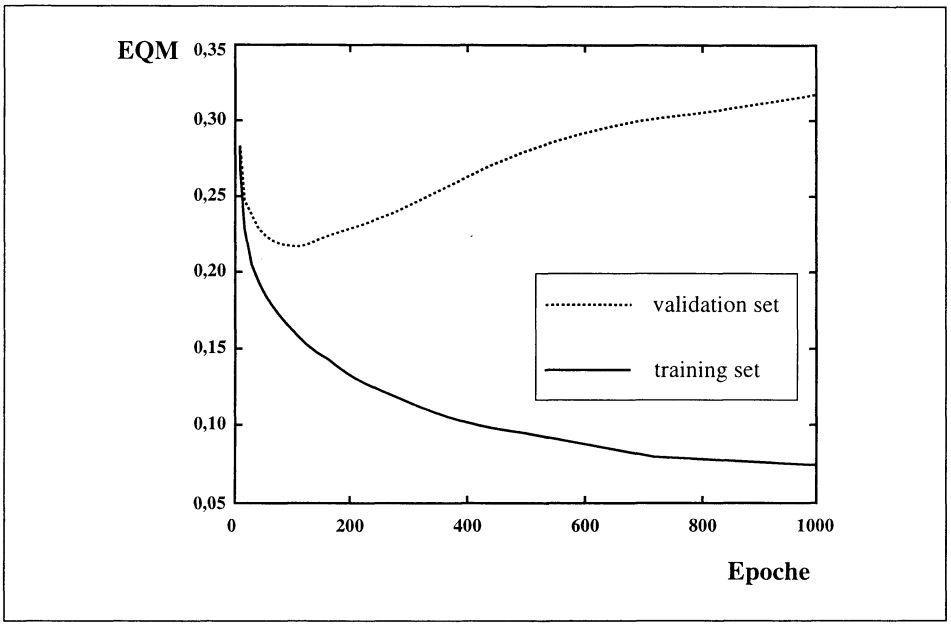

Fig. 7: Andamento tipico dell'errore sul training e sul validation set.

La presenza di un eventuale test set permette di effettuare un'ulteriore verifica della bontà del modello ottenuto. Il problema di tale tipo di tecnica è la forte dipendenza dalla particolare suddivisione del dataset iniziale nei tre sottoinsiemi, dipendenza tanto più forte quanto minore è la dimensionalità del campione a disposizione. Per ovviare a questo tipo di problema si è pensato di ricorrere alle tecniche di ricampionamento, tramite le quali è possibile, partendo dal campione

iniziale, generare differenti campioni sui cui addestrare e testare il modello. I metodi di ricampionamento sono differenti sia rispetto alla modalità con cui viene effettuata la generazione dei differenti campioni sia rispetto alle differenti accezioni di stabilità che intendono misurare. La stabilità del campione può cioè essere a sua volta suddivisa in due principali punti: *stabilità esterna* e stabilità *interna* (Davino *et al.*, 1997). Questa distinzione prende spunto dal contributo di Greenacre (1984) e distingue la stabilità in relazione alla presenza di osservazioni anomale nel campione (stabilità interna) da quella relativa alla rappresentatività dello stesso (stabilità esterna).

LA STABILITÀ INTERNA

La stabilità interna può essere misurata ricorrendo al jackknife o alla cross-validation (Stone, 1977), che ne rappresenta la naturale generalizzazione. Si tratta di tecniche che eliminano un gruppo di righe dalla matrice iniziale che utilizzano poi per testare il modello ottenuto sulla matrice ridotta ottenuta dalla depurazione delle stesse: lo scopo è verificare quanto gli elementi esclusi siano determinanti rispetto all'analisi. Il jackknife elimina una riga alla volta dalla matrice iniziale mentre la *cross validation* lavora su gruppi di righe. Si tratta di una generalizzazione del precedente metodo in quanto consiste nel dividere il campione iniziale di N osservazioni in un numero V, definito a priori, di sottoinsiemi disgiunti $£_1, £_2, \ldots, £_V$, ciascuno di ampiezza $d = N/V$, tali che $£_1 \cup £_2 \cup \ldots \cup £_V = £$ e $£_i \cap £_j = 0, \forall\, i \neq j$. Il metodo jackknife rappresenta quindi un caso particolare di cross validation in cui d=1, ovvero ciascun campione $£_i$ ha ampiezza 1.

La rete neuronale viene addestrata, fissata una inizializzazione dei pesi, utilizzando ciascuno dei V campioni $£ - £_v$, (v=1,...,V) come set di apprendimento ed i restanti $£_v$, (v=1,...,V) campioni esclusi per il test. Si hanno quindi a disposizione V percentuali di corretta classificazione sul training set, che possono essere utilizzate per la valutazione della stabilità interna, ed altrettante sul test set, che permettono invece di valutare la stabilità esterna. In Tabella 2 sono riportati i valori caratteristici delle distribuzioni delle percentuali di corretta classificazione nel caso di parametri $d = 10$ e $d = 20$. Il valore ottenuto sulla rete la cui stabilità è sottoposto all'analisi è pari a 83%: pur fornendo un valore apparentemente positivo, questo risulta esterno ai valori caratteristici risultanti dalle cross validation. Si può desumere quindi la presenza di una instabilità interna nei dati utilizzati per l'addestramento, il che può suggerire di valutare la rappresentatività del campione di dati a disposizione o della specifica suddivisione dello stesso in training set e test set utilizzata.

Tab. 2: Analisi della stabilità interna

	d = 10	*d = 20*
Minimo	78,89%	78,47 %
Media	79,60%	79,77 %
Massimo	81,11%	82,14 %
Scarto quadratico medio	0,31	0, 71
Coefficiente di variazione	0,0040	0,0089

LA STABILITÀ ESTERNA

Lo studio della stabilità esterna di una rete, ovvero della sua capacità di generalizzazione, può essere affrontato utilizzando il metodo bootstrap (Efron e Tibshirani, 1993). Sia $\mathbf{X}(N \times P)$ la matrice dei dati e $\mathbf{y}$ il vettore di output associato di dimensioni $(N \times 1)$: come è noto il *bootstrap* consiste nell'effettuare un campionamento con ripetizione da tale matrice in modo da estrarre un numero B definito a priori di matrici $\mathbf{X}^b$ (b=1,..,B) delle stesse dimensioni di quella iniziale con i corrispondenti vettori di output $\mathbf{y}^b$ (b=1,..,B).

La procedura di *bootstrap* è stata strutturata nei seguenti passi:

- Divisione del campione totale in un *test set* (T) ed in un *learning set* (L).
- Costruzione della regola di classificazione utilizzando l'insieme L ossia, nel contesto delle reti neuronali, addestramento della rete utilizzando l'insieme L come campione di apprendimento.
- Calcolo del tasso di errata classificazione sul test set (e(L,T)) sulla base dei pesi stimati al passo precedente.
- Generazione di B campioni L^b (b=1,..,B) bootstrap dall'insieme L e addestramento della rete su ciascuno di essi.
- Calcolo della percentuale di corretta classificazione utilizzando sempre il test set originario T, sulla base dei pesi ottenuti al passo precedente da ciascun campione bootstrap.

I risultati della fase di bootstrap con parametro B=100 sono rappresentati in Figura 8 unitamente agli estremi ([68,5%; 75,2%]) dell'intervallo di confidenza della distribuzione empirica centrato sulla media e di ampiezza pari ad uno scarto quadratico medio e comprendente circa il 68% della distribuzione dei valori. La percentuale di corretta classificazione ottenuta sulla rete oggetto di analisi è pari al 77%, valore che ricade all'esterno dell'intervallo empirico calcolato, comprovando quindi un eccessivo ottimismo nella valutazione della capacità di generalizzazione dei risultati ottenuti.

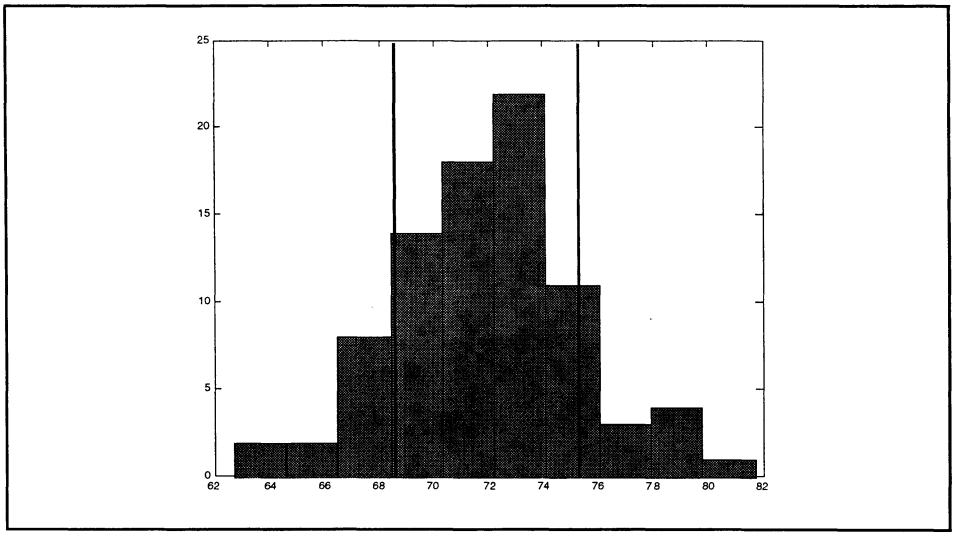

Fig. 8: Distribuzione delle % di corretta classificazione ottenute con il bootstrap.

L'approccio precedentemente illustrato consente di valutare a posteriori la stabilità delle previsioni ottenute; altre proposte di utilizzo dei metodi di ricampionamento si pongono invece l'obiettivo di ottenere delle previsioni già stabili in quanto risultato di una combinazione delle varie previsioni fornite dalle matrici ricampionate. Una delle varianti in questione è il recente *bagging* (*bootstrap aggregating*) (Breiman, 1994) (Breiman, 1996): l'idea è quella di ottenere differenti matrici di partenza ricampionando con ripetizione la matrice iniziale e valutando il modello su tali matrici.

Il *bagging* affronta sia il problema della scelta di uno specifico campione come set di apprendimento che il problema preliminare, ossia la divisione della matrice iniziale dei dati in un *test set* (T) ed in un *learning set* (L).

La procedura di *bagging* è strutturata nei seguenti passi:

- Divisione del campione totale in un *test set* (T) ed in un *learning set* (L).
- Costruzione della regola di classificazione utilizzando l'insieme L ossia, nel contesto delle reti neuronali, addestramento della rete utilizzando l'insieme L come campione di apprendimento.
- Calcolo del tasso di errata classificazione sul test set ($e(L,T)$) sulla base dei pesi stimati al passo precedente.
- Generazione di B campioni L^b (b =1,..,B) bootstrap dall'insieme L e addestramento della rete su ciascuno di essi.

- Calcolo della matrice degli output stimati sul test set ($O(L^b,T)$) sulla base dei pesi ottenuti al passo precedente da ciascun campione bootstrap.
- Aggregazione degli output stimati sul test set in modo da ottenere un unico vettore di sintesi da confrontare con l'output osservato sul test set. L'aggregazione viene realizzata, nel caso di una variabile di risposta numerica, sostituendo, per ciascun pattern, la media delle previsioni ottenute sui diversi campioni bootstrap e, nel caso di una variabile di risposta in classi, sostituendo, per ciascun pattern, la previsione prevalente ottenuta sui diversi campioni bootstrap.
- Calcolo del tasso di errata classificazione bagging ($e_b(L,T)$) confrontando l'output osservato sul test set con l'output aggregato al passo precedente.
- Ripetizione dei passi precedenti partendo da diverse divisioni del campione totale in *test set* e *learning set*.
- Calcolo dei tassi di errata classificazione medi $\bar{e}\,(L,T)$ e $\bar{e}_b(L,T)$.

Il tasso di errata classificazione bagging $\bar{e}\,(L,T) = 74{,}5\%$, con parametro B=100, risulta più basso del tasso di errata classificazione $\bar{e}\,(L,T) = 71\%$ ottenuto utilizzando una sola suddivisione del campione iniziale nell'insieme L e nell'insieme T garantendo, quindi, una migliore capacità di generalizzazione dei risultati della rete. La procedura di bagging, sebbene semplice ed efficace, presenta indiscussi costi computazionali: occorre addestrare B reti, conservare i pesi stimati e gli output stimati sul test set e, anche nella fase di previsione, occorre calcolare la previsione per il nuovo pattern utilizzando tutte le B reti addestrate.

4. CONSIDERAZIONI FINALI

L'utilizzo di una strategia integrata di analisi per le reti neuronali mette in luce come sia necessario un approccio critico all'applicazione di modelli che hanno come principale punto di forza la capacità di approssimare relazioni complesse, al fine di evitare risultati quanto meno discutibili rispetto ad uno o a più degli aspetti sopra presentati. L'esame di tali aspetti in un contesto applicativo ha messo in risalto come alcune scelte spesso scontate (ad esempio l'inizializzazione dei pesi o la suddivisione del campione in set differenti) portino a risultati fuorvianti; d'altro canto va tenuto comunque in considerazione l'alto costo computazionale richiesto dagli approcci basati su metodi di ricampionamento. L'oculatezza del ricercatore rimane quindi una tappa fondamentale per l'utilizzo delle reti, il cui contributo è fortemente influenzato dall'esperienza di chi le utilizza. La seguente citazione, suggerita in un contesto differente, ben descrive la necessaria sensibilità che si richiede all'utilizzatore di modelli neuronali, o più in generali di procedure automatiche di analisi:

"like any other tool, its greatest benefit lies in its intelligent and sensible application" (Breiman *et al.*, 1984).

Ulteriori sviluppi riguarderanno l'approfondimento di quegli aspetti della strategia (*pre-processing* e *post-processing* da un lato e *visualizzazione dei risultati* dall'altro) non trattati altrettanto dettagliatamente in questo lavoro. In particolare le fasi di pre-processing e post-processing possono essere analizzate in maniera incrociata, seguendo un approccio tipico dell'analisi di sensitività, ovvero andando a verificare l'effetto di piccole variazioni nelle scelte che caratterizzano queste fasi sulla performance finale del modello. Per quanto attiene la fase di visualizzazione dei risultati, oltre a poter sfruttare gli approcci tipici dell'analisi statistica, quali ad esempio un'analisi accurata dei residui, si potrebbe valutare la presenza di andamenti interessanti nella distribuzione dei pesi. A tal proposito ulteriori approfondimenti possono sicuramente interessare lo studio dell'incidenza dei pesi di cui al paragrafo 3.4.1, allo scopo di approfondire l'importanza relativa dei parametri nel modello.

BIBLIOGRAFIA

BISHOP C.M. (1995) – *Neural Networks for Pattern Recognition*. Oxford University Press.

BREIMAN L., J.H. FRIEDMAN, R.A. OLSHEN and C.J. STONE (1984) – *Classification and Regression Trees*. Wadsworth.

BREIMAN L. (1994). *Bagging Predictors*. Technical Report no. 421, University of California.

BREIMAN L. (1996). Bagging Predictors. *Machine Learning*, **26**, pp. 123-140.

BUCHANON B.G., E.H. SHORTLIFFE (1984) – *Rule-Based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison Wesley.

CHENG B., D.M. TITTERINGTON (1994) – Neural Networks: a Review from a Statistical Perspective. *Statistical Science, 9(1)*, pp. 1-54.

CIAMPI A., Y. LECHEVALLIER (1995) – Réseaux de neurones et modéles statistiques. *La revue de Modulad*, 15, pp. 27-46.

DAVINO C., GHERGHI M., VISTOCCO D. (1997) – Sulla Stabilità delle Regole di Classificazione nelle Reti Neuronali, in A. Bellacicco e C. Lauro (a cura di) *Reti Neurali e Statistica*, pp. 37-48, Franco Angeli.

EFRON B., R. J. TIBSHIRANI (1993) – *An introduction to the bootstrap*. Chapman & All.

GOLDBERG D.E. (1989) – *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley.

GREENACRE M.J. (1984) - *Theory and Applications of Correspondence Analysis*. Academic Press.

HAGAN M.T., H.B. DEMUTH E M.H. BEALE (1996) – *Neural Network Design*, Boston, MA: PWS Publishing.

HECHT-NIELSEN R. (1990) – *Neurocomputing*. Addison Wesley Pubblishing Company.

KOHONEN T. (1982) – Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, pp. 59-69.

LAURO C., DAVINO C., VISTOCCO D. (1999) – Neural Networks Applications in Economics: a Statistical Point of View, M. Marinaro, R. Tagliaferri (eds) *Neural Nets – Proceedings of the 11th Italian Workshop on Neural Nets*, pp. 357-375, Springer.

MCCULLOCH W.S., W. PITTS (1943) – A Logical Calculus of Ideas Immanent in Neural Activity. *Bulletin of Mathematical Biophysics*, 5.

MOODY J. (1994) – Prediction Risk and Architecture Selection for Neural Networks. In V. Cherkassky, J.H. Friedman, H. Wechsler (eds.), *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*, pp. 146-165.

RIPLEY B.D. (1994) – Neural Networks and Related Methods for Classification, *Journal of the Royal Statistical Society*, 56(3), pp. 409-456

RIZZI A. (1998) – Algoritmi Genetici e Cluster Analysis. *METRON – International Journal of Statistics.*

RUMELHART D., G. HINTON, R. WILLIAMS (1986) – Learning Represesentations by Backpropagating Errors. *Nature*, 323, pp. 459-473.

SARLE W.S. (1995) – Stopped Training and Other Remedies for Overfitting, *SAS Institute.*

SARLE W.S. (1996) – Neural Networks and Statistical Jargon. *SAS Institute*, ftp://ftp.sas.com/pub/neural/jargon.

SMITH M. (199) – *Neural Networks for Statistical Modelling*. International Thomson Publishing.

STEVENS L. (1984) – *Artificial Intelligence. The Search for the Project Machine.* Hayden Book Company, Hasbrouck Heights.

STONE M. (1977) – An asymtotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society*, B39, pp. 44-47.

TIBSHIRANI R. (1996) – *A comparison of some error estimates for neural networks models.* Technical Reports, ftp://utstat.toronto.edu/pub/tibs/nnerr.ps.

ZADEH L.A. (1977) – Fuzzy Sets and Their Application to Pattern Classification and Clustering, J.VAN RYZIN (ed.), *Classification and Clustering*, Academic Press.

NEURAL NETWORKS STABILITY IN AN INTEGRATED STATISTICAL STRATEGY

Summary

Statistical applications of neural networks are recently widespread. The aim of this paper is to propose an integrated statistical strategy to use neural networks in a statistical framework. Particular attention is addressed to the stability of the model from the point of view both of the parameters stability and of the sample stability. The sample stability is investigated distinguishing an internal and an external stability thus considering the presence of anomaulus observations and the representativeness of the sample. The architecture selection, the definition of the model and the training phases are faced using different simulations before choosing the "best" one in order to take into account simultaneously performance and parsimony goals.