Università di Firenze
Facoltà di Psicologia

Dario Aversa
Rosapia Lauro Grotto

Modelli di Reti neurali

www.aracneeditrice.it
info@aracneeditrice.it

via Raffaele Garofalo, 133 A/B
00173 Roma
(06) 93781065

ISBN 978–88–548–2016–6

I edizione: settembre 2008

a Mariano Arcelloni

1. Origine delle reti neurali

Lo studio delle reti neurali risale ai primi tentativi di tradurre in modelli matematici i principi dell'elaborazione neurale biologica. Le più antiche teorie del cervello e dei processi mentali sono state concepite dai filosofi greci Platone (427–347 a.C.) e Aristotele (384–322 a.C.). Queste teorie furono riprese molto più tardi da Cartesio (1586–1650) e nel XVIII secolo dai filosofi empiristi. Le prime realizzazioni di *macchine cibernetiche*, categoria alla quale appartengono i sistemi neurali, appaiono negli anni quaranta col nascere di una scienza nuova, la ***cibernetica***. La cibernetica è definita come scienza che studia i processi intelligenti e fu fondata da Norbert Wiener nel 1947. Ross Ashby, un altro padre della cibernetica, costruisce nel 1948 l'*omeostato* uno dei primi sistemi con connessioni interne regolabili, capace quindi di variare la sua configurazione interna adattandola a stimoli esterni. Il neurofisiologo W.S. McCullock e il matematico W.A. Pitts (1943) di Chicago sono stati i primi a formulare l'approccio cibernetico fondamentale alla struttura del cervello elaborando il primo modello di rete neurale. John Von Neumann, dopo aver formulato l'architettura base dei moderni calcolatori, comincia nel 1948 lo studio delle *reti di automi cellulari* precursori di nuovi modelli computazionali. Nel 1949 il neurofisiologo Donald Hebb, dagli studi sul processo di apprendimento dei neuroni, dedusse la prima *regola di apprendimento* applicata nelle reti neurali (Hebb, 1949). La “regola di Hebb” prevede che

se due neuroni collegati fra loro sono attivi contemporaneamente il valore sinaptico delle loro connessione viene aumentato. Contemporaneamente gli studi di Lashley (Lashley, 1950) sulla mente umana indicavano che l'organizzazione della conoscenza e della memoria si basava su *rappresentazioni distribuite,* in cui ogni elemento è rappresentato da un *pattern* di attività distribuito su molte unità di computazione, e ogni unità è usata per rappresentare molti elementi differenti (McClelland e Rumelhart, 1986; trad. it. 1991). Il punto di forza di questo schema di rappresentazione è l'efficacia con cui sfrutta la capacità di elaborazione propria delle reti formate da unità di computazione elementari simili a neuroni.

Nei primi anni Sessanta si costruiscono le prime macchine in grado di presentare primitive forme di apprendimento spontaneo e guidato, sono il ***Perceptron*** di Frank Rosemblatt della *Cornell University* e l'***Adaline*** *(Adaptive Linear Element)* di Bernard Windrow di Stanford. *Il Perceptron è una rete neurale costituita da dispositivi logici in grado di risolvere semplici problemi di riconoscimento di forme.* Esso rappresentò un prototipo delle strutture che vennero elaborate più avanti. *L'Adaline o "Regola Delta" è un algoritmo di apprendimento basato su un procedimento iterativo di correzione dell'errore molto potente nella capacità di generalizzare a nuovi pattern*.

Anche in Italia si sviluppano iniziative particolarmente importanti. Eduardo Caianello, dell'Università di Napoli, sviluppa la sua teoria sui processi e le macchine pensanti sulla base delle

idee di Mc Cullock, Pitts e Hebb. A Genova viene realizzata da *Augusto Gamba* una macchina derivata dal *Perceptron*.

Nel 1969 Marvin Minsky e Seymour Papert, del *Massachusetts Institute of Technology*, pubblicano un'analisi molto critica delle macchine tipo il *Perceptron*. Nel loro volume *Perceptrons* essi dimostrarono matematicamente le limitazioni delle reti neurali nel risolvere funzioni logiche che non erano linearmente separabili, come la funzione XOR [f(0,0) = f(1,1) = 0, f(1,0) = f(0,1) = 1]. Questi problemi potevano essere risolti solo da reti neurali omniconnesse in cui ogni neurone è connesso con tutti gli altri neuroni della rete; in una simile rete il numero delle connessioni crescerebbe esponenzialmente all'aumentare del numero di neuroni contrariamente a quanto avviene nei sistemi biologici nei quali le connessioni crescono linearmente (Minsky e Papert, 1969). Minsky era uno dei sostenitori di un approccio rivale alle reti neurali, l'***Intelligenza Artificiale (A.I.) classica*** basata su computer tradizionali. In seguito alle tesi di Minsky il campo delle reti neurali fu abbandonato dalla maggior parte degli studiosi, i quali si rivolsero al campo dell'A.I. apparentemente più promettente. Questo cambiamento di interessi fu causato anche dal fatto che la tecnologia allora disponibile rendeva molto difficoltosa o addirittura impossibile la sperimentazione nel campo delle reti neurali, né vi erano computer abbastanza veloci per simulare reti neurali complesse. Negli anni sessanta e settanta la ricerca continuò con contributi teorici e poche applicazioni. Alcuni ricercatori come Shunichi Amari, Kunihiko Fukushima e

Shephen Grossberg tentarono di simulare il comportamento di neuroni cerebrali con reti di unità di calcolo operanti in modalità parallela.

L'interesse sviluppatosi nei primi anni ‘80 per i modelli neurali è sicuramente dovuto a diversi fattori, che sono elencati di seguito:

— i progressi compiuti nella comprensione di alcuni fenomeni computazionali biologici, la disponibilità di potenti computer in grado di simulare i nuovi modelli neurali.

— la determinazione dei limiti dell'A.I., i quali sono strettamente legati ai limiti dei computer seriali di Von Neumann.

John Hopfield del *California Institute of Technology* propone nel 1982 un modello computazionale basato su concetti energetici e pertanto applicabile in svariati campi. Nel 1986 Dave Rumelhart e Jay McClelland pubblicarono due volumi su modelli a *processamento distribuito in parallelo* (*Parallel Distributed Processing:* PDP), oggi considerati in qualche modo la “bibbia” del connessionismo, e con cui si vide la diffusione dell’algoritmo di *Back–propagation,* “scoperto” più volte e già prima da altri autori (Bryson & Ho, 1969; Le Cun, 1985; Parker, 1985; Werbos, 1974). Questo algoritmo propone un potente metodo ricorsivo per modificare i pesi sinaptici di una rete neurale con un qualsiasi numero di strati composti a loro volta da un qualsiasi numero di neuroni.

Con questo risultato termina la "preistoria" dello studio delle reti neurali e inizia la cronaca di un settore in rapida evoluzione.

2. Definizione teorica di una rete neurale artificiale

Le **reti neurali artificiali** sono dei sistemi di elaborazione dell'informazione il cui funzionamento trae ispirazione dai sistemi nervosi biologici (Floreano, 1996).

Esamineremo quindi brevemente alcune delle caratteristiche principali del Sistema Nervoso biologico umano.

L'unità computazionale costitutiva di tutto il sistema nervoso è il **neurone**. Il neurone, a cui si ispirano i modelli di reti neurali, è una cellula dotata di corpo cellulare dal quale dipartono molte brevi ramificazioni d'ingresso (*dendriti*) e una sola lunga ramificazione di uscita (*assone*). Il cervello umano è costituito da una rete di moltissimi neuroni ($10^{11} \div 10^{12}$) collegati tra loro da un numero elevato di connessioni dette **sinapsi** ($10^{3} \div 10^{4}$ sinapsi per ogni neurone). In totale il numero di sinapsi del cervello varia da 10^{14} a 10^{16} (Kandel et al., 1998) . Esse regolano la quantità di segnale e il tipo di effetto che esso provoca sul neurone ricevente. Ogni sinapsi può essere inibitoria o eccitatoria per cui il numero delle configurazioni possibili è:

$$N = (2^{10})^{12}$$

I neuroni comunicano tra loro mediante impulsi particolari detti *spike* o *potenziali d'azione*, generati da processi elettrochimici. Il livello di tensione a cui la membrana del neurone dà origine allo *spike* è detto soglia di eccitazione. Lo *spike* si propaga lungo l'assone liberando energie locali e sulla sinapsi provoca l'emissione di particolari sostanze dette *neurotrasmettitori*,

i quali raggiungono i dendriti del neurone successivo e provocano una variazione della permeabilità della membrana, consentendo la trasmissione dei segnali tra i neuroni connessi. L'effetto additivo di più potenziali d'azione sui dendriti del neurone postsinaptico capace di superare la soglia d'attivazione del neurone genera un potenziale d'azione. Il numero di potenziali d'azione nell'unità di tempo in funzione della corrente in ingresso va da 0 a 100 ÷ 1000 *spike* per secondo (Kandel et al., 1998).

L'elaborazione dell'informazione nel sistema nervoso avviene in parallelo, e questo spiega l'incredibile velocità del cervello nell'eseguire compiti che richiedono l'elaborazione contemporanea di un elevato numero di dati. Inoltre vi sono molti neuroni che si occupano della stessa operazione e quindi l'elaborazione è distribuita su molti elementi. Questo fa sì che il cervello sia un sistema estremamente flessibile a superare disastri locali nella sua struttura senza perdita significativa di prestazioni (Ladavas & Berti, 1999).

Le osservazioni fatte in relazione alla computazione biologica permettono di affermare che è possibile definire i principi costruttivi di nuove macchine di calcolo capaci di superare i limiti degli elaboratori tradizionali. Questi nuovi sistemi di elaborazione dell'informazione sono chiamati reti neurali. Una rete neurale artificiale è composta da molte unità (neuroni, nodi o processori) di elaborazione variamente connesse fra di loro. Ogni unità diviene attiva se la quantità totale di segnale che riceve supera

la propria soglia di attivazione. L'unità che si attiva emette un segnale che giunge fino alle altre unità a cui è connessa. I punti di connessione (sinapsi o pesi sinaptici) sono dei filtri che trasformano il segnale in eccitatorio o inibitorio, aumentandone o diminuendone l'intensità. Il segnale di risposta di un nodo è funzione della somma dei prodotti dei segnali d'ingresso per i rispettivi pesi sinaptici, meno il valore della soglia del nodo. Quando uno stimolo (vettore o *pattern* in *input*) viene applicato ai neuroni d'ingresso (*input)* della rete, i segnali viaggiano in parallelo lungo le connessioni attraverso i nodi interni (*hidden*) fino ai nodi di uscita (*output*) la cui attivazione rappresenta la risposta della rete neurale.

La configurazione delle connessioni (architettura) e i valori dei pesi sinaptici determinano in gran parte il comportamento e la risposta della rete.

Una rete impara a fornire le risposte appropriate per ciascuno stimolo in ingresso modificando i valori delle proprie connessioni sinaptiche in base a regole di apprendimento, che prescindono dal tipo di compito per cui la rete verrà utilizzata.

Le principali caratteristiche delle reti neurali sono:

Robustezza: una rete è resistente al danneggiamento, ovvero è in grado di continuare a dare una risposta sostanzialmente corretta anche se alcune delle sue connessioni vengono eliminate.

Flessibilità: un modello neurale può essere impiegato per un gran numero di finalità diverse: esso non ha bisogno di cono-

scere le proprietà del dominio specifico di applicazione perché apprende in base all'esperienza.

Capacità di generalizzazione: una rete neurale che è stata addestrata su un numero limitato di esempi è in grado di produrre una risposta adeguata, senza più modificare le connessioni sinaptiche, a pattern d'ingresso che non ha mai visto in precedenza ma che presentano tuttavia qualche somiglianza con gli esempi presentati durante la fase di apprendimento. Uno dei problemi che si può incontrare durante la generalizzazione è l'*overfitting*, ossia quando un gran numero di pesi e unità permette alla rete di apprendere una vasta serie di funzioni specifiche che sono responsabili della corrispondenza esatta tra input e output per i pattern di addestramento, diminuendo così la probabilità che la rete scopra proprio la funzione generale che descrive l'intero dominio del problema e che darebbe risposte corrette anche per i pattern di test (Floreano, 1996).

Capacità di elaborare segnali rumorosi (resistenza al rumore): le reti neurali sono in grado di recuperare le proprie memorie in base al contenuto partendo da dati incompleti, simili o corrotti dal rumore.

Un numero troppo grande di connessioni compromette anche la convergenza verso un minimo globale, sia perché aumenta la complessità della superficie d'errore, sia perché richiede un alto numero di *pattern* d'addestramento (Baum e Haussler, 1989). La scelta del numero di *hidden* e di pesi sinaptici è quindi un problema cruciale.