

Non ti avevo riconosciuto...

Analisi dei dati e ricostruzione di segnali

di Tommaso Boccali

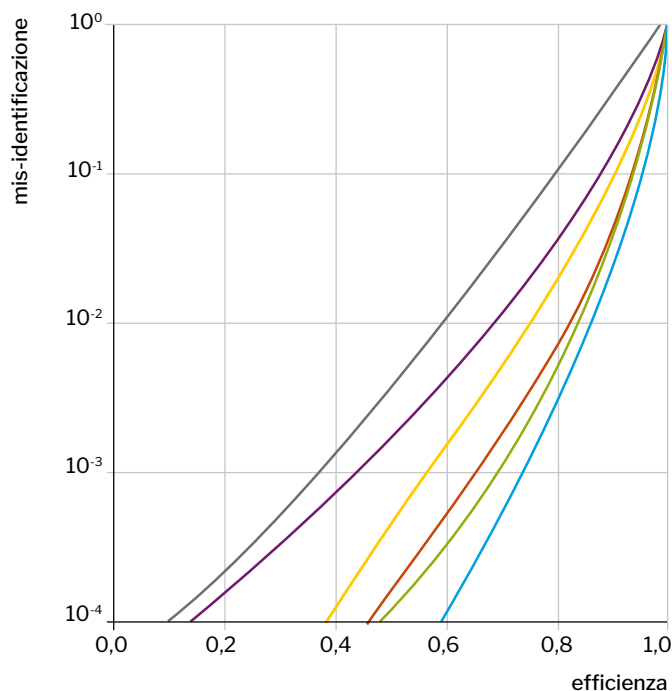


a.
L'esperimento ALEPH del Large Electron-Positron Collider (LEP), l'acceleratore di particelle del CERN di Ginevra negli anni '90 (precursore di LHC). L'utilizzo delle reti neurali per la ricerca di rari decadimenti del mesone B nell'esperimento ALEPH può essere considerato il primo esempio di uso dell'intelligenza artificiale nella fisica delle alte energie.

Il 30 novembre 2022 ChatGPT 3.5 è stato reso disponibile mediante ogni browser web, provocando una presa di coscienza da parte del grande pubblico delle potenzialità dell'intelligenza artificiale.

Nel mondo della ricerca, e della ricerca in fisica fondamentale in particolare, metodi ispirati al funzionamento del cervello umano sono utilizzati da un tempo molto maggiore. È ovviamente difficile risalire a un primo utilizzo, ma possiamo comunque considerare come uno dei primi nella fisica delle alte energie quello descritto nell'articolo di Philippe Rosnet e Pierre Henrard del 1997, sull'utilizzo delle reti neurali per la ricerca di rari decadimenti del mesone B nell'esperimento ALEPH del LEP, in cui viene proposta una struttura a 20 neuroni di input e due strati nascosti, con rispettivamente 15 e 10 neuroni. Il risultato viene condensato in un solo neurone nell'ultimo strato, per un

totale di 486 “pesi”, usati per calibrare la rete neurale. Una rete con tali caratteristiche era, a quei tempi, al limite della tecnologia informatica utilizzabile. Ai giorni nostri, l'utilizzo di grandi infrastrutture di calcolo e di processori dedicati come le GPU permette di addestrare e utilizzare reti con centinaia di miliardi di “pesi”. Insieme a una maggiore comprensione teorica delle reti neurali (in buona parte dovuta a studi che considerano tali reti come “sistemi complessi”, vd. p. 10, ndr) e alla disponibilità di modelli più avanzati, questa esplosione delle capacità tecnologiche è la maggior responsabile della “rivoluzione” IA nella fisica fondamentale. Oggi metodologie IA sono usate in quasi tutti gli ambiti della ricerca in tale campo: dalla simulazione degli eventi, capace di rivaleggiare con simulazioni dettagliate che hanno rappresentato lo standard fino ad oggi, alla ricostruzione delle particelle partendo dai segnali elettronici



— Likelihood ratio
 — Multi Layer Perceptron
 — Deep Neutral Network
 — Convolutional Neutral Network + Recursive Neutral Network
 — Dynamic Graph Convolutional Neutral Network
 — Transformer

b.

Un tipico grafico “mis-identificazione vs efficienza”, che mostra il miglioramento delle prestazioni introdotte dall’intelligenza artificiale nella fisica delle alte energie. La separazione tra eventi di segnale e rumore è tanto migliore quanto più è bassa la mis-identificazione (incorretta identificazione di eventi di fondo) a parità di efficienza nella corretta identificazione di eventi di segnale. Si nota come, man mano che si passa da algoritmi semplici e privi di IA (come Likelihood ratio) ad algoritmi che usano IA avanzata (p.es. i Transformer), la mis-identificazione si riduce sempre di più a parità di efficienza. L’obiettivo è avvicinarsi il più possibile all’angolo in basso a destra del diagramma.

grezzi, alle tecniche di analisi dei dati. Focalizzandoci su quest’ultima categoria, ci sono due esempi che possono essere utili per capire l’impatto dell’intelligenza artificiale nel settore. Uno dei segnali più interessanti nella fisica ai collisori ad alte energie degli ultimi decenni è rappresentato dagli eventi con la presenza di quark b (bottom, o beauty), il secondo quark più pesante dei sei del modello standard delle particelle elementari. La rivelazione di tali quark è un indicatore importante per moltissimi canali di interesse: la fisica dei sapori pesanti, dell’Higgs, del top e in moltissimi scenari oltre il modello standard. Tali quark, prodotti in misura minore rispetto a quark più leggeri, generano adroni con caratteristiche peculiari come una maggiore massa, una maggiore vita media e decadimenti con un numero elevato di particelle. Nessuna di queste caratteristiche è tale da permettere una separazione perfetta dagli eventi di rumore, per cui l’identificazione di eventi originati da quark b è fatta su base statistica. Le prestazioni degli algoritmi di IA si basa su due parametri: la loro efficienza nella corretta identificazione di eventi di segnale, in questo caso i quark b (detta “efficienza”), e alla percentuale di incorretta identificazione di eventi che non li contengono, il cosiddetto “fondo” (detta “mis-identificazione”). A parità di efficienza, l’algoritmo funziona meglio se la mis-identificazione è bassa.

Queste prestazioni possono essere riportate su un grafico bidimensionale come quello di fig. b, in cui l’algoritmo migliore è rappresentato dalla regione in basso a destra. Questo grafico ci mostra anche l’evoluzione di tali algoritmi di selezione per i quark b, che esistono da decenni. Tali algoritmi si sono evoluti passando da semplici selezioni basate su una sola delle caratteristiche sopra elencate a combinazioni sempre più complesse di diverse informazioni. Con l’arrivo dell’intelligenza artificiale, inizialmente con metodi come

Boosted Decision Trees e poi con semplici reti neurali dense (i cosiddetti Multi Layer Perceptron), che combinano un certo numero di caratteristiche in modo ottimizzato, si ottengono discriminazioni sempre migliori.

Arrivano poi tecniche avanzate di IA, come il *deep learning* e reti basate su grafi e i Transformer, che combinano un numero molto alto di caratteristiche, comprese alcune che a prima vista non paiono eccessivamente rilevanti, ma per le quali l’apprendimento automatico riesce a trovare comunque correlazioni tra le variabili utili, anche se piccole. Il risultato migliore ottenuto ad oggi, utilizzando metodi di intelligenza artificiale avanzati è, a parità di efficienza, più di 50 volte migliore del migliore degli algoritmi vecchio stile. Nella pratica questo vuol dire che un’analisi di fisica avrà 50 volte meno fondo e sarà in grado di selezionare segnali molto più puri. In aggiunta, in molti casi, c’è anche la richiesta di avere più quark b nell’evento, anche 2, 3 o 4. In questo caso il guadagno sulla mis-identificazione del fondo può scalare con le potenze di 50 (una per ogni quark b aggiuntivo), e quindi essere molto rilevante ai fini della precisione delle misure. Sono numeri ancora più significativi, con un impatto notevole che ha consentito la realizzazione di misure di fisica estremamente migliori di quanto inizialmente prospettato.

Il secondo esempio riguarda la rivelazione delle onde gravitazionali, uno dei campi di ricerca di più grande interesse, passato dalla fase di scoperta all’essere un efficace strumento per la ricerca astrofisica e cosmologica. I segnali raccolti in esperimenti come Virgo sono molto flebili, al limite della sensibilità sperimentale. In tale regime, la capacità di discriminazione fra segnali indotti da rumori locali legati all’ambiente esterno o ai sottosistemi del rivelatore e segnali reali è essenziale, ancor di più in un’ottica di ricerca multimessaggera, in cui l’allerta sulla



C.
Vista aerea
dell'interferometro
di onde gravitazionali
Virgo, a Cascina (nei
pressi di Pisa).

rivelazione dei segnali gravitazionali deve essere veloce per permettere ad altri esperimenti di orientarsi per una rivelazione congiunta. La pronta classificazione di segnali sperimentali (tipicamente mostrati in un diagramma tempo-frequenza detto “spettrogramma”) come appartenenti a classi di effetti indotti dal rivelatore o da eventi esterni (i cosiddetti “*glitch*”) è essenziale in questo processo ed è un’attività complessa, che non può essere svolta con metodi classici e algoritmici. E l’IA riesce a svolgere un ruolo cruciale anche in questa classificazione. Dal momento che i dati su cui si opera sono immagini (diagrammi tempo-frequenza), l’uso delle reti neurali convoluzionali (CNN) sulle immagini degli spettrogrammi è ideale. Perché l’algoritmo funzioni è necessario un set di immagini con la giusta etichetta che un umano esperto oppure un altro algoritmo possono fornire: evidentemente la qualità finale dell’algoritmo CNN dipende dalla qualità e dalla quantità di tale set. Una possibile soluzione per avere a disposizione

tale set di immagini è appoggiarsi a soluzioni di *citizen science*, nelle quali viene chiesto a una comunità di esperti ed entusiasti di aiutare in tale classificazione, ai fini della preparazione di un set da usare come allenamento per l’algoritmo. I risultati sono molto buoni e il tempo di esecuzione della classificazione è dell’ordine di pochi microsecondi, ottimale per caratterizzare i segnali di *glitch* in modo da studiarne l’origine per poterli rimuovere e poter quindi usare i segnali reali di onde gravitazionali per inviare rapidamente eventuali *alert* da sottoporre agli altri sistemi di rivelazione multimessaggera. Quelli descritti sono solo due esempi di come l’intelligenza artificiale possa essere utile nell’analisi dei dati della fisica fondamentale. In realtà, oggi è raro trovare ambiti di ricerca in cui l’IA non venga già usata o il cui uso non sia previsto per il prossimo futuro. I ricercatori dell’INFN, che lavorano in vari campi scientifici, stanno imparando a conoscere e usare queste tecnologie in modo efficace, così da far progredire la scienza.

Biografia

Tommaso Boccali è dirigente di ricerca presso l’INFN di Pisa. Già durante il suo dottorato presso la Scuola Normale Superiore, si è occupato di problematiche di calcolo scientifico, inizialmente per l’esperimento ALEPH e poi per l’esperimento CMS a LHC. Attualmente è supervisore del progetto CNAF Reloaded, che prevede l’integrazione delle risorse INFN al CNAF presso il Tecnopolo di Bologna, e *spoke 2 leader* del Centro Nazionale Big Data, HPC e Quantum Computing.