theFuture ofScience andEthics

Rivista scientifica a cura del Comitato Etico
della Fondazione Umberto Veronesi

Volume 2 **numero 2** ■ dicembre 2017

Fondazione Umberto Veronesi
– per il progresso delle scienze

theFuture ofScience andEthics

Fondazione Umberto Veronesi
– per il progresso delle scienze

theFuture ofScience andEthics

Rivista scientifica
del Comitato Etico
della Fondazione Umberto Veronesi
ISSN 2421-3039
ethics.journal@fondazioneveronesi.it
Periodicità semestrale
Piazza Velasca, 5
20122, Milano

Direttore
Cinzia Caporale

Condirettore
Silvia Veronesi

Direttore responsabile
Donatella Barus

Comitato Scientifico
Roberto Andorno (University of Zurich, CH); Massimo Cacciari (Università Vita-Salute San Raffaele, Milano); Stefano Canestrari (Università di Bologna); Carlo Casonato (Università degli Studi di Trento); Roberto Cingolani (Direttore scientifico Istituto Italiano di Tecnologia-IIT, Genova); Giancarlo Comi (Direttore scientifico Istituto di Neurologia Sperimentale, IRCCS Ospedale San Raffale, Milano); Gilberto Corbellini (Sapienza Università di Roma e Consiglio Nazionale delle Ricerche-CNR); Lorenzo d'Avack (Università degli Studi Roma Tre); Giacinto della Cananea (Università degli Studi di Roma Tor Vergata); Sergio Della Sala (The University of Edinburgh, UK); Hugo Tristram Engelhardt jr. (Rice University e Baylor College of Medicine, Houston, TX, USA); Andrea Fagiolini (Università degli Studi di Siena); Daniele Fanelli (London School of Economics and Political Science, UK); Gilda Ferrando (Università degli Studi di Genova); Giovanni Maria Flick (Presidente emerito della Corte costituzionale); Nicole Foeger (Austrian Agency for Research Integrity-Oe-AWI, Vienna, e Presidente European Network for Research Integrity Offices — ENRIO); Tommaso Edoardo Frosini (Università degli Studi Suor Orsola Benincasa, Napoli); Filippo Giordano (Libera Università Maria Ss. Assunta-LUMSA, Roma); Giorgio Giovannetti (Rai — Radiotelevisione Italiana S.p.A.); Massimo Inguscio (Presidente del Consiglio Nazionale delle Ricerche CNR); Giuseppe Ippolito (Direttore scientifico IRCCS Istituto Nazionale per le Malattie Infettive Lazzaro Spallanzani, Roma); Michèle Leduc (Directrice de recherche émérite au CNRS et Comité d'étique du CNRS, FR); Luciano Maiani (Sapienza Università di Roma e CERN, CH); Sebastiano Maffettone (LUISS Guido Carli, Roma); Elena Mancini (Consiglio Nazionale delle Ricerche-CNR); Vito Mancuso (Teologo e scrittore); Alberto Martinelli (Università degli Studi di Milano); Roberto Mordacci (Università Vita-Salute San Raffaele, Milano); Paola Muti (McMaster University, Hamilton, Canada); Ilja Richard Pavone (Consiglio Nazionale delle Ricerche-CNR); Renzo Piano (Senatore a vita); Alberto Piazza (Università degli Studi di Torino e Presidente dell'Accademia delle Scienze di Torino); Riccardo Pietrabissa (Politecnico di Milano); Tullio Pozzan (Università degli Studi di Padova e Consiglio Nazionale delle Ricerche-CNR); Francesco Profumo (Politecnico di Torino e Presidente Fondazione Bruno Kessler, Trento);

Giovanni Rezza (Istituto Superiore di Sanità-ISS); Gianni Riotta (Princeton University, NJ, USA); Carla Ida Ripamonti (Fondazione IRCCS Istituto Nazionale dei Tumori-INT, Milano); Angela Santoni (Sapienza Università di Roma); Pasqualino Santori (Presidente Comitato Bioetico per la Veterinaria-CBV, Roma); Elisabetta Sirgiovanni (Sapienza Università di Roma e New York University); Guido Tabellini (Università Commerciale Luigi Bocconi, Milano); Henk Ten Have (Duquesne University, Pittsburgh, PA, USA); Giuseppe Testa (Istituto Europeo di Oncologia-IEO, Milano); Chiara Tonelli (Università degli Studi di Milano); Silvia Veronesi (Avvocato); Riccardo Viale (Scuola Nazionale dell'Amministrazione-SNA e Herbert Simon Society); Luigi Zecca (Consiglio Nazionale delle Ricerche-CNR).

Sono componenti di diritto del Comitato Scientifico della rivista i componenti del Comitato Etico della Fondazione Umberto Veronesi: Cinzia Caporale (Presidente del Comitato Etico) (Consiglio Nazionale delle Ricerche-CNR); Vittorino Andreoli (Psichiatra e scrittore); Elisabetta Belloni (Segretario Generale Ministero degli Affari Esteri e della Cooperazione Internazionale); Gherardo Colombo (già Magistrato della Repubblica italiana, Presidente Casa Editrice Garzanti, Milano); Carla Collicelli (Consiglio Nazionale delle Ricerche-CNR); Domenico De Masi (Sapienza Università di Roma); Giuseppe Ferraro (Università degli Studi di Napoli Federico II); Carlo Flamigni (Comitato Nazionale per la Bioetica); Vittorio Andrea Guardamagna (Istituto Europeo di Oncologia-IEO); Antonio Gullo (Università degli Studi di Messina); Armando Massarenti (CNR Ethics); Lucio Militerni (Consigliere emerito Corte Suprema di Cassazione); Telmo Pievani (Università degli Studi di Padova); Carlo Alberto Redi (Università degli Studi di Pavia e Accademia Nazionale dei Lincei); Alfonso Maria Rossi Brigante (Presidente onorario della Corte dei conti); Marcelo Sánchez Sorondo (Cancelliere Pontificia Accademia delle Scienze); Paola Severino Di Benedetto (Rettore LUISS Guido Carli, Roma); Elena Tremoli (Università degli Studi di Milano e Direttore scientifico IRCCS Centro Cardiologico Monzino, Milano).

Coordinatore del Comitato Scientifico: Laura Pellegrini

Redazione: Marco Annoni (Caporedattore) (Consiglio Nazionale delle Ricerche-CNR); Giorgia Adamo (Consiglio Nazionale delle Ricerche-CNR); Chiara Mannelli (Università di Torino, Candiolo Cancer Institute, FPO - IRCCS); Annamaria Parola (Fondazione Umberto Veronesi); Roberta Martina Zagarella (Consiglio Nazionale delle Ricerche-CNR).

Progetto grafico: Gloria Pedotti

Volume 2 **numero 2** ■ Dicembre 2017*

SOMMARIO

ARTICOLI

CALL FOR PAPERS: CURABILI E INCURABILI

DOCUMENTI DI ETICA E BIOETICA

RECENSIONI

NEWS *a cura di Giorgia Adamo*

theFuture ofScience andEthics

** Il numero è stato chiuso il 31 gennaio 2018.*

Call for papers: "Curabili e incurabili"

Intelligenza artificiale, *Machine learning* e *Big Data*: concetti di base e applicazioni nelle bioscienze

Artificial Intelligence, Machine Learning and Big Data: basic principles and bioscience applications

PAOLA BERTOLAZZI
paola.bertolazzi@iasi.cnr.it

AFFILIAZIONE
SYSBIO.IT Center for Systems Biology,
Università degli Studi Milano Bicocca e
Consiglio Nazionale delle Ricerche

ABSTRACT

Intelligenza artificiale, *Machine Learning* e *Big Data* sono tra gli argomenti più caldi che compaiono quasi ogni giorno sui media in occasione della presentazione di molti risultati della ricerca. I primi due sono stati studiati fin dagli anni cinquanta, mentre i *Big Data* sono un concetto che appare di recente (2011) e che indica le nuove tecnologie in grado di gestire i dati che sono distribuiti sul web in grande dimensione e con formati diversi. Mentre l'intelligenza artificiale è un'area molto ampia che include la robotica, la dimostrazione di teoremi, la comprensione del linguaggio naturale, i sistemi esperti e altri argomenti, l'apprendimento automatico è uno dei temi della intelligenza artificiale e riguarda i metodi che conferiscono a un programma di computer la capacità dell'essere umano e animale di apprendere da esempi per acquisire la capacità di riconoscere situazioni o prevedere tendenze future. L'apprendimento automatico è un argomento cruciale, in quanto i *Big Data* richiedono di essere analizzati per estrarre conoscenze che nessun essere umano potrebbe ottenere in altro modo. La bioscienza è un campo in cui questi due argomenti giocano un ruolo centrale a causa della quantità di dati che vengono generati quotidianamente dalle moderne tecnologie genomiche e dalla complessità dei sistemi biologici che richiedono l'indagine della relazione tra un grande numero di elementi.

KEYWORDS

Intelligenza Artificiale
Artificial Intelligence

Apprendimento Automatico
Machine Learning

Big Data
Big Data

Bioscienze
Bioscience

ABSTRACT

Artificial Intelligence, Machine Learning and Big Data are among the hotter topics that appear almost every day when research results are presented to common people by media. The first two have been investigated since the 50s, while Big Data is a concept that appeared recently (2011) to include new technologies to manage data that are distributed on the web and are not represented in usual formats. While AI is a very large area that include robotics, theorem proving, natural language comprehension, expert systems and other arguments, machine learning is one of the AI subjects and concerns the methods that can provide a computer program with the ability of human and animal to learn from examples and acquire the ability to recognize situations or predict future trends. Machine learning is now a crucial topic, since Big Data require to be analyzed to extract knowledge that no human could obtain in other ways. Bioscience is a field where these two topics play a central role due to the amount of data that are daily generated by modern genomic technologies and the complexity of biological systems which require the investigation of relation among a very large number of elements.

Sempre più, la stampa e altri mezzi di comunicazione riportano e commentano notizie circa potenzialità e risultati dell'Intelligenza Artificiale, del *Machine Learning* e dei *Big Data* nello sviluppo di soluzioni a problemi che emergono in numerosi settori della realtà economica e sociale, ivi compreso quello biomedico.

Un esempio è quello della notizia diffusa recentemente da molti quotidiani circa la possibilità di diagnosticare la patologia di Alzheimer con dieci anni di anticipo grazie alle tecniche di Intelligenza Artificiale e Machine Learning utilizzate da un gruppo di scienziati di Bari per interpretare le immagini di Risonanza Magnetica dei cervelli di pazienti affetti da questa patologia[1] (Rasero 2017). Altri esempi si possono trovare in lavori pubblicati su questa stessa rivista (Riotta 2016; Scalzini 2016; De Maldè 2017).

Le due aree disciplinari sopra citate nascono intorno alla metà degli anni Cinquanta. Entrambe sono fondate su teorie matematiche e tecnologie informatiche il cui sviluppo prosegue incessantemente per affrontare sempre nuovi problemi di rappresentazione della realtà e di complessità del calcolo[2]. Oggi, una delle sfide più importanti per queste discipline è il mondo dei *Big Data* che le mette significativamente alla prova in termini di gestione e analisi di queste immense quantità di informazioni.ù

1. L' INTELLIGENZA ARTIFICIALE

L'Intelligenza Artificiale viene fondata come area disciplinare nel 1956, durante un *workshop* di due mesi svoltosi presso Dartmouth College, da un gruppo di dieci scienziati provenienti dal Massachusetts Institute of Technology (MIT), Carnegie Mellon University (CMU), IBM e altri centri di ricerca. In quell'occasione ne viene coniata la denominazione e ne vengono fissati i principi identificando non solo le problematiche del futuro ma anche i riferimenti bibliografici in cui si possono riconoscere gli albori dell'area, dai testi di Omero fino alla letteratura *fantasy* e fantascientifica. Anche le macchine che erano state realizzate prima del primo calcolatore, fra cui quelle per il gioco degli scacchi, vengono annoverate come primi esempi di *thinking* artificiale[3,4] (Buchanan 2006).

In questa prima fase significativi passi avanti vengono fatti nella dimostrazione automatica di teoremi e nella capacità da parte del computer di esprimersi in lingua inglese.

Nel 1974 la ricerca sulla IA si ferma per poi riprendere, negli anni Ottanta, con lo studio dei Sistemi Esperti, sistemi informatici che sono costituiti da una base di formule logiche (Base di Conoscenza) costruite a partire dalle competenze di esperti, che vengono elaborate con algoritmi di *Reasoning* (gli stessi che venivano usati per la dimostrazione dei teoremi) e permettono di risolvere molti problemi in diversi settori fra i quali uno dei principali è quello della salute, dove supportano il medico sia nella diagnostica sia nella scelta delle cure (De Maldè 2017). Anche in questo caso però la ricerca, dopo aver toccato un picco in termini di finanziamenti nel 1985, viene sostanzialmente abbandonata nel 1987. Verso la fine degli anni Novanta, il successo del computer *Deep Blue*, che riesce a battere il giocatore di scacchi Garry Gasparov, riporta l'IA agli onori della cronaca: in quel momento metodologie e tecnologie sono tali da rendere più promettenti i risultati di questa disciplina.

Diversi sono i problemi che vengono fatti ricadere nell'area della IA, e che hanno caratterizzato le diverse fasi dello sviluppo della disciplina: la dimostrazione di teoremi, i sistemi esperti, i giochi, la comprensione del linguaggio naturale, la robotica, il *data mining*, il riconoscimento delle immagini e il *machine learning.* Ognuna di queste applicazioni richiede che il sistema informativo a essa dedicato abbia alcune delle seguenti capacità: *reasoning*, *problem solving*, rappresentazione della conoscenza, pianificazione, apprendimento, percezione, movimento, comprensione del linguaggio.

Ad esempio, nella robotica sono necessari software per la percezione (in termini di visione artificiale e sensoristica) e la pianificazione che sono tipici della IA, mentre la parte di attuazione (movimento e manipolazione) viene effettuata da software basati principalmente sulla matematica che descrive i sistemi di controllo e permette di modellare in modo digitale la catena fra la percezione e l'attuazione del gesto. Nei sistemi esperti e di supporto alle decisioni, sono importanti sia metodi e tecniche per la rappresentazione della conoscenza, per memorizzare le competenze degli esperti in forma di regole logiche, sia algoritmi per il *reasoning/problem solving.*

Secondo la definizione di Russel (2010), in tutti i casi sopra elencati, l'approccio IA consiste nello studio di agenti che ricevono percetti (oggetti della percezione) dall'ambiente e effettuano azioni, intendendosi per studio la modellazione di tali agenti e loro realizzazione come Applicazioni Software, le quali, simulando le capacità intellettive umane e animali, siano in grado di rispondere a determinati stimoli con azioni adeguate. Deve quindi essere progettato un modello computabile di tali agenti che realizzi la funzione del cervello desiderata e deve esistere un calcolatore che sia in grado di eseguire questo modello.

Il modello può replicare in modo abbastanza fedele il funzionamento della mente umana, oppure non corrispondere affatto al comportamento del cervello, ma seguire un procedimento completamente diverso, che giunge tuttavia allo stesso risultato. Al primo tipo di modelli appartengono quelli utilizzati nei sistemi di supporto alle decisioni in medicina, che sono costituiti da regole logiche, prodotte dall'esperto, del tipo "se il paziente ha la febbre alta e le placche in gola" allora "ha una infezione da batterio". Quando il medico deve produrre una diagnosi inserisce nell'applicazione una serie di informazioni sul paziente e il sistema di *reasoning* associa i dati immessi alle formule logiche verificando se queste risultino vere o false in relazione ad essi. Il processo seguito è simile a quello umano. Un altro esempio è quello dei sistemi in grado di giocare a scacchi che si basano su procedimenti di calcolo che

esaminano le migliaia di possibili mosse di una partita al fine di scegliere quella da attuare in un determinato momento del gioco. Al secondo tipo appartengono invece i modelli che vengono utilizzati per l'analisi e riconoscimento di immagini.

Da questa brevissima trattazione si è visto come l'IA copra numerose tematiche, che continuano a essere oggetto di sviluppi incessanti. Tuttavia gli aspetti che, almeno all'apparenza, stanno prendendo il sopravvento su tutto il resto, in questa ultima fase della storia della IA, sono quelli dell'apprendimento, riconoscimento e previsione, a cui ci si riferisce generalmente con il termine di *Machine Learning*. Questo interesse è principalmente dovuto al sempre crescente accumularsi di ogni genere di dati, fenomeno che è stato chiamato qualche anno fa "data deluge", e che attualmente ha preso il nome di *Big Data*. Questi due temi saranno trattati nelle prossime sezioni.

2. IL MACHINE LEARNING

Il termine viene coniato alla fine degli anni Cinquanta da Arthur Samuel di IBM, esperto di teoria dei giochi, uno dei dieci scienziati che avevano contribuito a fondare, pochi anni prima, l'intelligenza artificiale[5] (Samuel 1959). Il *Machine Learning* (Mitchell 1997; Kohavi 1998) studia metodi e tecnologie che permettono da una parte di simulare quell'aspetto dell'intelligenza che riguarda la capacità dell'apprendimento attraverso esempi, applicando a tal fine le teorie del pattern recognition e dell'apprendimento computazionale, dall'altra di utilizzare ciò che è stato appreso per riconoscere e predire.

La visione o riconoscimento d'immagini rappresenta uno dei primi esempi di applicazione del *Machine Learning.* L'apprendimento avviene dando in *input* al computer tante immagini di uno stesso oggetto d'interesse; il computer esegue un software che è in grado di estrarre delle caratteristiche sintetiche (*feature*) comuni a queste immagini. Questo software apprende il concetto nel senso che costruisce un modello sintetico dell'oggetto, basato sulle *feature* e sulle relazioni fra di esse. Un secondo software, dipendente in qualche modo dal primo, ricevendo in *input* un'immagine di un oggetto simile, sarà in grado di cercare in questa immagine quelle stesse caratteristiche e una volta trovatele potrà dichiarare di avere riconosciuto nell'immagine l'oggetto.

Questo complesso di concetti (*feature*, sequenza di esempi) e di architettura di calcolo (un software per l'estrazione de*l pattern*, un altro per il riconoscimento) viene utilizzato con dati di input molto diversi. Oltre alle immagini, i dati che maggiormente vengono trattati in questo modo sono tabelle di valori, che possono essere numerici (valori interi o reali), logici o nominali.

Una tabella, o matrice, di valori appare come un elenco di righe di valori, incolonnati fra loro. Ogni riga corrisponde a un esempio (individuo) della realtà che stiamo studiando e ogni colonna viene associata a una *feature*. Un esempio di tabella che possiamo facilmente immaginare è una tabella ISTAT sulle cui righe troviamo tutti i cittadini italiani e sulle colonne le informazioni raccolte con il censimento. Nell'ambito della salute, sulle righe potremmo trovare un insieme di individui affetti da una certa patologia e sulle colonne informazioni legate all'età, condizione sociale, zona geografica, dati clinici e dati biologici.

Il lavoro che un sistema di *Machine Learning* dovrà fare sulla tabella sarà lo stesso che abbiamo descritto nel caso delle immagini. La tabella corrisponde a un elenco di esempi che verranno usati per l'apprendimento. Tali esempi possono essere o non essere già raggruppati in classi. Nel caso lo siano, l'apprendimento viene detto *supervisionato*, e consiste nel cercare delle *feature* comuni a ciascuna di tali classi e che differenzia gli esempi di questa classe rispetto a tutte le altre classi. Tali caratteristiche comuni rappresentano il modello della classe in base al quale si effettuerà il riconoscimento. Nel caso gli esempi non siano raggruppati in classi, l'apprendimento è invece detto *non supervisionato*, e avviene attraverso la ricerca di gruppi di esempi che abbiano caratteristiche comuni, dai quali verrà estratto il modello che verrà usato per il riconoscimento.

Prima di approfondire le questioni sulla natura dei modelli che caratterizzano i vari gruppi di esempi, estratti dal software di apprendimento, forniamo un breve cenno su come avviene l'apprendimento nel caso '*supervisionato*'. Immaginiamo che la tabella, le cui righe sono associate a diversi esempi appartenenti a due classi diverse, che chiameremo per comodità A e B, sia divisa in due insiemi di esempi, il primo detto di *training* ed il secondo detto di *test*, entrambi contenenti elementi delle due classi. L'insieme degli esempi di *training* viene usato per identificare i modelli che caratterizzano le due

classi. La validità dei modelli trovati viene sottoposta a *test* usando l'altro sottoinsieme di esempi la cui classe di appartenenza non viene segnalata al software di riconoscimento. La capacità di riconoscere correttamente la classe di appartenenza degli esempi nel test sulla base dell'osservazione degli esempi nell'insieme di *training* è il principale indicatore della capacità di apprendimento del sistema.

Tornando ai modelli, questi si distinguono fondamentalmente in due classi: modelli con semantica e modelli senza semantica (che potremmo definire scatola nera). I modelli con semantica sono tipicamente costituiti da formule logiche costruite su un sottoinsieme delle informazioni contenute nelle colonne della tabella. Fra i metodi che sono in grado di estrarre dalle tabelle questo tipi di modelli richiamiamo il *data mining* logico (Boros 1997; Felici 2002; Truemper 2004; Felici 2005) con recenti estensioni (Fiscon 2016; Cestarelli 2016) e gli alberi di decisione (Rokach 2008).

I modelli senza semantica possono essere formule matematiche o addirittura algoritmi che non mettono in luce il ruolo giocato dalle informazioni contenute nelle colonne della tabella. Fra questi ricordiamo le *support vector machine* (Boser 1992) e le reti neurali (McCulloch 1943).

Il tipo di modello che si utilizza nel software di apprendimento / riconoscimento / previsione viene deciso dal progettista del software che spesso fa diverse prove utilizzando svariati tipi di modelli per valutarne la qualità in relazione alle capacità di riconoscimento del software stesso e delle caratteristiche dei dati analizzati. Un esempio interessante di modelli a formule logiche sono quelli che emergono nella classificazione di specie animali o di virus, che caratterizzano ciascuna specie con una formula che indentifica quali valori debbano avere alcune posizioni nel DNA (Weitschek 2013; Bertolazzi 2015).

3. I BIG DATA

Con questo termine, ancora non ben stabilizzato, vengono identificate le problematiche connesse alla gestione e alla fruizione della massa di dati destrutturati che si vanno accumulando a causa della diffusione imponente di tutte le tecnologie digitali che producono informazioni. Il termine è stato coniato nel 2011 e va a sostituire altri termini fra i quali quello di "*data deluge*" (diluvio dei dati) che rappresentava in modo figurativo il fenomeno che si andava delineando sempre più chiaramente. Da allora si sono andati chiarendo i confini di quest'area, che tratta fondamentalmente lo sviluppo e uso di tecnologie per la memorizzazione, gestione e elaborazione di queste masse di informazioni[6].

Gli aspetti che caratterizzano i *Big Data* sono la destrutturazione e l'eterogeneità dei dati, che fanno sì che tutte le tecnologie per la gestione delle Basi di Dati divengano inutilizzabili. Nelle Basi di Dati tradizionali i dati sono organizzati in tabelle, come quelle sopra descritte, memorizzate in grandi banche centralizzate, con formati definiti e coerenti. I *Big Data* invece si accumulano ovunque, e sono memorizzati con formati diversi. Poiché il problema principale nella gestione dei dati è quello di analizzarli per poterli mettere in relazione fra loro, cosa tutto sommato facile da realizzare nel caso dei dati contenuti in una Base di Dati tradizionale, la sfida straordinaria nei *Big Data* è quella di realizzare tecnologie che permettano di mettere in relazione dati che, come già detto, sono fra loro eterogenei (perché provenienti da fonti eterogenee), rappresentati con formati diversi (da quelli dei dati strutturati, come i database, a quelli non strutturati, come immagini, email, dati GPS, informazioni prese dai *social network*) ed infine distribuiti in memorie situate in giro nella rete.

Quelle che oggi vengono presentate come le tecnologie per i *Big Data* riguardano essenzialmente due aspetti: come memorizzare e gestire questi dati distribuiti ovunque e come analizzarli.

Per il primo aspetto esistono fondamentalmente due soluzioni, prodotte dalle due delle più importanti società di software esistenti, MapReduce di Google e Hadoop di Apache, che fondamentalmente risolvono i seguenti problemi: gestire file distribuiti su computer qualsiasi per permettere un'elaborazione "parallela" su porzioni di questi dati, opportunamente ritagliate, e condurre le elaborazioni di queste porzioni a un'ultima fase di calcolo che ricomponga i risultati parziali per fornire il risultato finale.

Il secondo aspetto sembra possa essere affrontato con una tecnica generale di *Machine Learning,* detta *Deep Learning* introdotta in Rumelhart (1986). Tale metodo si ispira alle reti neurali, ovvero metodo a scatola nera per il *Machine Learning*, basato su una tecnica di apprendimento che consiste nell'assegnare ai nodi di una rete che simulano dei neuroni dei

pesi che si individuano attraverso l'esame di esempi da apprendere. Alla fine dell'addestramento, i pesi dei nodi della rete saranno tali da far seguire a un nuovo esempio la cui classe sia ignota un percorso che lo faccia giungere a un nodo finale il quale ne indichi la classe di appartenenza. Nel *Deep Learning* (LeCun 2015) le reti neurali che vengono costruite sono a molti livelli e questo consente l'uso delle nuove tecnologie hardware basate sul paradigma del calcolo parallelo, che comporta velocità di calcolo incomparabilmente più alte del passato. Tale metodo viene usato principalmente nel riconoscimento di immagini e nella comprensione del linguaggio naturale.

Nella prossima sezione faremo un cenno ai problemi connessi con l'analisi dei dati che vengono prodotti nell'area delle bioscienze e in particolare della biologia molecolare.

4. LE APPLICAZIONI DEL MACHINE LEARNING ALLE BIOSCIENZE: OPPORTUNITÀ E LIMITI

Il *Machine Learning* viene attualmente impiegato massicciamente per l'analisi di dati biologici che vengono prodotti in quantità sempre maggiori, anche attraverso progetti che finanziano la raccolta di dati secondo protocolli stabiliti a livello globale e che permettono di ottenere dati estratti da esperimenti fatti nelle stesse condizioni. Come si comprende, l'impatto potenziale di questi studi è altamente innovativo e significativo per le singole persone e per la socità.

Ci riferiamo in questa sezione alle problematiche connesse all'esame di particolari dati che, dall'avvento delle tecnologie e metodologie per il sequenziamento del DNA, vengono prodotti in quantità sempre crescenti e a costi sempre più ridotti. Il *Next Generation Sequencing* (NGS), una delle tecniche maggiormente impiegate, permette di ricavare la sequenza completa del DNA di un individuo e tutte le informazioni relative (per esempio la presenza di certe mutazioni), e può determinare la quantità di RNA (espressione) presente in questo tessuto, sia per quanto riguarda le porzioni codificanti (geni) del DNA sia per quanto riguarda le porzioni non codificanti. Tecnologie più precise permettono anche di misurare la quantità di proteine o di metaboliti.

Queste informazioni possono essere tradotte in tabelle di valori e sottoposte ad analisi del tipo di quelle sopra descritte. La maggior parte delle analisi che vengono effettuate oggi riguarda i dati di espressione di RNA codificante (espressione genica) e puntano a individuare processi biologici disfunzionanti in individui affetti da patologie (Arisi 2011; Arisi 2015). La sfida è però quella di analizzare tutti i dati prodotti da un esperimento di NGS in maniera integrata. In questo caso la complessità della computazione diviene insostenibile ed è impossibile pensare di attuarla su calcolatori tradizionali anche se molto potenti. Basti pensare che mentre le informazioni sull'espressione genica raggiungono le decine di migliaia, quelle sulle mutazioni sono centinaia di migliaia.

Si possono presentare, a titolo esemplificativo, tre situazioni in cui gli approcci sopra descritti non sono sufficienti per estrarre le informazioni desiderate.

Un primo caso riguarda la ricerca nelle malattie genetiche. Nel passato l'approccio seguito era tendenzialmente quello di esaminare un gene alla volta per verificare se certe mutazioni avessero un legame con certe patologie. Risalgono agli anni Sessanta, ben prima del sequenziamento dell'intero DNA umano, i primi risultati su correlazioni fra singoli tratti del DNA e malattie genetiche (McKusick 1969; 1988; 2001). Ben presto ci si è resi conto che per molte patologie non erano riscontrate tali semplici correlazioni ma si poteva ipotizzare che la causa della malattia fosse il mal funzionamento di due o più geni contemporaneamente. In tali casi si deve ricorrere a un approccio, detto poligenico, che prevede quindi la ricerca di anomalie concomitanti in più di un gene. Se si dovesse procedere, con metodi puramente algoritmici, alla valutazione di tutte le possibili combinazioni di mutazioni, poiché il numero di possibili siti di mutazione oggi noti raggiunge i due milioni di unità, il tempo di calcolo necessario per completare tali valutazioni esploderebbe in maniera combinatoria senza portare ad alcun risultato in tempi ragionevoli.

Gli approcci oggi considerati risolutivi per l'analisi dei *Big Data* a poco servono in questi casi. Come abbiamo detto in precedenza, infatti, le reti neurali sono tecniche che offrono un modello senza semantica, quindi inutilizzabile se vogliamo sapere quali mutazioni di quali geni siano la possibile ragione di una malattia. Esistono delle versioni parallele degli alberi di

decisione, che riescono a analizzare, in tempi ragionevoli, fino a 450000 siti, ma certamente passare a dimensioni superiori richiede un miglioramento drastico degli algoritmi.

Un altro caso in cui tali tecniche sono poco utilizzabili è quello in cui non sono mutazioni di geni a essere la causa di una malattia, bensì il malfunzionamento di processi biologici. I geni contribuiscono al funzionamento della vita attraverso la produzione di proteine che interagiscono fra loro, con l'aiuto di altre molecole (RNA, zuccheri, etc.), in reazioni a catena che vengono rappresentate come piccole reti (o "grafi") e che vengono chiamate *pathway* o processi biologici. Lo studio delle semplici tabelle ottenibili dai dati prodotti dal NGS non può fornire molte informazioni sul malfunzionamento di questi processi, ma occorre individuare nuovi metodi che permettano di integrare le informazioni sulle reazioni note o estrarre informazioni sull'esistenza di reazioni non ancora studiate. In particolare è necessario utilizzare l'approccio della *Systems Biology* (Snoep 2005) che permette di tenere conto in vari modi della complessità dei legami che esistono fra le varie molecole.

Sono oggi note migliaia di relazioni fra coppie di molecole, rappresentate in maniera compatta con le semplici reti di proteine (PPI) o in maniera più espressiva nei data base di processi biologici, ma la ricerca di nuove relazioni non si ferma. Sarà poi necessario studiare modelli matematici di questi processi biologici per poter utilizzare tali modelli per simulare situazioni di patologie o di cura di queste patologie, riducendo così, ad esempio, il numero degli esperimenti su animali cavie o rendendo più efficiente la sperimentazione sugli umani.

La terza problematica riguarda la scarsa confrontabilità di dati raccolti con protocolli diversi. Accade spesso che, a fronte di dati sulla stessa patologia, raccolti in ricerche diverse, i metodi di analisi più
usati diano luogo a soluzioni molto diverse in termini di geni e di processi mal funzionanti. Questo potrebbe voler significare che non sarà possibile pensare di incrociare dati *bio* presi da internet senza rischiare di ottenere risultati privi di senso.

CONCLUSIONI

Quelle che sono state presentate sono le attuali tecnologie e metodologie con le quali si pensa di poter raggiungere una migliore comprensione di moltissimi problemi la cui complessità è tale da non poter essere capita e dominata con i modelli attualmente noti e trattabili. La speranza che i dati possano suggerire relazioni che nessun ricercatore o gruppo di ricercatori può riuscire a individuare senza questi strumenti va coltivata con grande prudenza.

A poco servono calcolatori sempre più potenti. La complessità dei calcoli che devono essere eseguiti cresce esponenzialmente con il volume dei dati, e non sarà quindi gestibile nemmeno con la parallelizzazione del calcolo (Cook 1980). Altresì, a poco serve accumulare terabyte di dati se questi non vengono armonizzati e resi confrontabili.

Per quanto riguarda in particolare le bioscienze e la biomedicina, ciò che occorre è un maggiore impegno del mondo della ricerca allo scopo di sviluppare nuovi modelli e algoritmi che siano in grado di rappresentare in modo compatto la complessità dei sistemi a cui sono associati questi dati. Occorre inoltre un sempre maggiore sforzo per sostenere la ricerca interdisciplinare affinché l'affiancamento di matematici, statistici, fisici, ingegneri, informatici biologi e ricercatori biomedici possa permettere una sempre maggiore comprensione dei fenomeni oggetto di studio, quali i meccanismi di sviluppo delle malattie e quelli del funzionamento dei farmaci.

Molto negli Stati Uniti ed in Europa è stato fatto da questo punto di vista: esistono centri di ricerca, più o meno specializzati in particolari malattie, dove molte delle competenze sopra indicate lavorano spesso all'interno di centri di cura e dove si riesce ad attuare quella che viene definita medicina traslazionale. Così come, per quanto riguarda i dati, esistono numerosissime iniziative di raccolta di dati in forma standardizzata, relative a singole patologie e su DNA[9].Purtroppo le esperienze europee rispetto alla ricerca interdisciplinare restano iniziative lasciate alle scelte dei singoli Stati e l'Italia dove, soprattutto a livello della ricerca pubblica, prevale il consueto e autolesionistico atteggiamento accademico di chiusura o pretesa di far prevalere una disciplina rispetto alle altre, i pochi centri esistenti, quasi tutti di natura privata, difettano nelle dimensioni e nell'entità dei finanziamenti.

La politica della Commissione europea sui finanziamenti alla ricerca rimane ancora settorializzata mentre, se si considerano le scelte fatte sulle Infrastrutture di Ricerca, l'unico esempio che coltiva l'ambizione di

creare un ambiente europeo interdisciplinare per il supporto allo sviluppo di nuovi modelli, è la nuova infrastruttura di *Systems Biology* (ISBE), che ancora deve completare la sua *roadmap*. Molte delle altre iniziative approvate riguardano reti di laboratori, fino ad ora sostenuti dai singoli Stati nazionali, che raccolgono tecnologie hardware e software per la memorizzazione e gestione di dati e immagini (come ELIXIR o Euro-BioImaging). Sarà loro compito provvedere agli aspetti di armonizzazione dei dati per permettere una sempre maggiore possibilità di integrazione di dati di natura diversa. Ma senza la presenza di un'infrastruttura per lo sviluppo di modelli questi dati rischiano di rimanere infruttuosamente sequestrati nella miniera.

Infine un cenno sulle questioni etiche: questo argomento è esaurientemente trattato nel lavoro di Scalzini del 2016, dove sono identificate le problematiche prioritarie nel settore della salute, che vanno ben al di là delle questioni della privacy e della anonimia, ma toccano principalmente l'uso che istituzioni, datori di lavoro o assicurazioni sanitarie potrebbero fare dei risultati delle analisi dei dati stessi a fini di previsione e riconoscimento, nelle decisioni sui dipendenti o sugli assicurati. Un tema non nuovissimo in bioetica, ma che gli sviluppi tecnologici e la potenza di calcolo rilanciano con molta forza. Anche sul piano etico, tuttavia, sarà essenziale la capacità di aggregare scelte e procedure a livello internazionale, in modo da facilitare lo sviluppo di una ricerca eticamente sostenibile.

NOTE

1. Si vedano i lavori riguardanti questo tema al link https://scholar.google.it/citations?user=J99kNm4AAAAJ&hl=it

2. Sulla differenza fra queste due are e l'area della Scienza dei Dati (Data Science) si veda http://varianceexplained.org/r/ds-ml-ai/

3. Cfr. http://digitalcollections.library.cmu.edu/awweb/awarchive?type=file&item=38698

4. Cfr. https://en.wikipedia.org/wiki/Artificial_intelligence

5. Cfr. https://en.wikipedia.org/wiki/Machine_learning

6. I riferimenti alle principali soluzioni tecnologiche possono essere rinvenuti al seguente link molto semplice e chiaro: https://it.wikipedia.org/wiki/Big_data

7. Cfr. https://cancergenome.nih.gov

8. Al riguardo si veda TCGA - The Cancer Genome Atlas degli NIH (https://cancergenome.nih.gov/), o ADNI su immagini di cervelli da pazienti con Alzheimer (http://adni.loni.usc.edu).

9. Gli NIH in USA ospitano una delle più imponenti raccolte di dati genetici sulle mutazioni del DNA. Cfr. https://www.ncbi.nlm.nih.gov/probe/docs/projhapmap/

BIBLIOGRAFIA

• Arisi I., D'Onofrio M., Brandi R, Felsani A, Capsoni S, Drovandi G., Giovanni Felici, Emanuel Weitschek, Paola Bertolazzi, Cattaneo A: Gene Expression Biomarkers in the Brain of a Mouse Model for Alzheimer's Disease: Mining of Microarray Data by Logic Classification and Feature Selection., Journal of Alzheimer's Disease 24, 2011.

• Arisi I., D'Onofrio M., Brandi R., Cattaneo A., Paola Bertolazzi, Fabio Cumbo, Giovanni Felici, Concettina Guerra: Time dynamics of protein complexes in the AD11 transgenic mouse model for Alzheimer's disease like pathology, Bmc Neuroscience, 2015.

• Bengio Yoshua, 2009, Learning Deep Architectures for AI (PDF), in Foundations and Trends in Machine Learning, vol. 2.

• Bertolazzi, P., Felici, G., Fiscon, G., Weitschek, E. (2015): Classifying DNA barcode multi-locus sequences with feature vectors and supervised approaches, Genome 58(5).

• Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A. (1997). A Logical Analysis of Numerical Data. Mathematical Programming.

• Boser, B. E.; Guyon, I. M.; Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers". Proceedings of the fifth annual workshop on Computational learning theory – COLT '92. p. 144.

• Buchanan, B. G. (2006) A (Very) Brief History ofArtificial Intelligence, AI Magazine Volume 26 Number 4 (© AAAI).

• Cestarelli, V., Fiscon, G., Felici, G., Bertolazzi, P., Weitschek, E. (2016): CAMUR: Knowledge extraction from RNA- seq cancer data through equivalent classification rules. Bioinformatics, 32(5).

• Cook S.A. 1980 Towards a complexity theory of synchronous parallel computation presented at Internales Symposium über Logik und Algorithmik, Enseign. Math., 27, Zurich.

• De Maldè, M. (2017), Medicina di precisione e sistemi di supporto alla decisione clinica: opportunità di miglioramento delle cure, riduzione degli errori e contenimento dei costi, The Future of Science and Ethics vol. 2 n. 1.

• Felici, G., Truemper K., (2002) A Minsat Approach for Learning in Logic Domains, INFORMS Journal Of Computing 14 (1).

• Felici, G., Truemper, K. (2005), The Lsquare System for Mining Logic Data, Encyclopedia of Data Warehousing and Mining.

• Fiscon, G., Weitschek, E., Cella E., Lo Presti, A., Giovanetti, M. Babakir-Mina, M., Ciotti, M., Ciccozzi, M., Pierangeli, A., Bertolazzi, P., Felici, G. (2016): MISSEL: a method to identify a large number of small species-specific genomic subsequences and its application to viruses classification, BioData Mining.

• Kohavi, R. and Provost, F. (1998), Glossary of terms," Machine Learning, vol. 30, no. 2-3.

• LeCun, Y., Bengio, Y. and Hinton, G. E. 2015, Deep Learning. Nature, Vol. 521, pp 436-444.

• McCulloch, Warren; Walter Pitts (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biophysics. 5 (4): 115–133. doi:10.1007/BF02478259.

• McKusick, Victor (1969). "On lumpers and splitters, or the nosology of genetic disease". Perspect Biol Med. 12 (2): 298–312. doi:10.1353/pbm.1969.0039. PMID 4304823.

• McKusick, Victor (1988). "Probable Assignment of the Duffy Blood Group Locus to Chromosome 1 in Man". Proc. Natl. Acad. Sci. U.S.A. 61 (3): 949–55. doi:10.1073/pnas.61.3.949. PMC 305420 Freely accessible. PMID 5246559.

• McKusick, Victor (2001). "The Anatomy of the Human Genome: a Neo-Vesalian Basis for Medicine in the 21st Century". The Journal of the American Medical Association. 286 (18): 2289–95. doi:10.1001/jama.286.18.2289. PMID 11710895.

• Mitchell, T. (1997). Machine Learning. McGraw Hill.

• Rasero Javier, Amoroso Nicola, La Rocca Marianna, Sabina Tangaro, Roberto Bellotti, Sebastiano Stramaglia, (2017) Multivariate regression analysis of structural MRI connectivity matrices in Alzheimer's disease PLoS ONE 12(11): e0187281.

• Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc.

• Riotta, Gianni (2016), Big Data, Sed Data. L'era degli algoritmi, dal potere dei dati al mistero della narrativa. The Future of Science and Ethics vol. 1 n. 2.

• Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (1986). "Learning representations by back-propagating errors". Nature. 323 (6088): 533-6.

• Russel Stuart, Norvig Peter (2010) Artificial Intelligence - A Modern Approach III Ed., Prentice Hall.

• Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development.

• Scalzini, Silvia (2016), Big Data e integrità della ricerca: un punto di partenza, The Future of Science and Ethics vol. 1 n. 2.

• Snoep, Jacky L; Westerhoff, Hans V (2005). Alberghina, Lilia; Westerhoff, Hans V, eds. "Systems Biology: Definitions and Perspectives". Topics in Current Genetics. Topics in Current Genetics. Berlin: Springer-Verlag. 13: 13–30. doi:10.1007/b106456. ISBN 978-3-540-22968-1.

• Truemper K. (2004), Design of Logic-Based Intelligent Systems, Wiley-Interscience.

• Weitschek, E., Van Velzen R, Felici, G., Bertolazzi, P. (2013): BLOG 2.0: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it, Molecular Ecology Resources.

• Weitschek, E., Lo Presti, A., Drovandi, G., Felici, G., Ciccozzi, M. Ciotti, M., Human polyomaviruses identification by logic mining techniques, Virology journal 9 (1), 58.