Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Informatica per il Management

ALCUNI CASI DI RISULTATI FUORVIANTI NELL'APPLICAZIONE DEL MACHINE LEARNING IN MEDICINA

Relatore:
Chiar.mo Prof.
MARCO ROCCETTI

Presentata da:
RICCARDO MOLINARI

Sessione II
2017/2018

Introduzione

Ormai da anni si assiste alla progressiva introduzione ed adozione dell'Intelligenza Artificiale in ogni settore lavorativo, economico e nella vita di ciascun individuo.
In alcuni casi è talmente integrata nella nostra quotidianità che risulta difficile accorgersi della sua presenza ed influenza.
Anche le tecniche di Machine Learning hanno fatto progressi sostanziali in molti settori dell'industria, come ad esempio in Medicina dove è addirittura considerata la grande speranza del 21esimo secolo per migliorare le prospettive di vita dell'umanità.
In questo campo ha già portato grandi trasformazioni nel sistema sanitario e si presume che porterà ulteriori progressi nei prossimi anni: fornirà ai medici un supporto sempre più sicuro ed efficiente nel raccogliere, organizzare, analizzare i dati clinici, fare diagnosi precoci e trovare migliori soluzioni per i pazienti.
Esistono previsioni per cui l'apprendimento automatico determinerà cambiamenti sostanziali nell'assistenza sanitaria, in particolare nelle discipline che richiedono modelli prognostici più precisi, come in oncologia, e quelli basati sul riconoscimento patologico, come in radiologia.
Sebbene l'introduzione dell'utilizzo di questi algoritmi abbia portato miglioramenti e benefici in Medicina, non esiste ancora la certezza che un loro utilizzo massivo e generalizzato garantisca un incremento nell'efficacia dell'attività clinica. Infatti il risultato delle predizioni degli algoritmi può talvolta essere fuorviante e controverso, portando il medico a decisioni sbagliate.

In generale, nella storia della Medicina, l'introduzione di nuove tecnologie non è mai stata semplice e senza effetti indesiderati, proprio per la natura stessa di questa scienza non deterministica.

Indice

Elenco delle figure

Capitolo 1

Intelligenza Artificiale

1.1 Cenni storici

Nell'ultimo secolo l'uomo ha cercato di costruire macchine che potessero aiutarlo nel calcolo e nelle proprie attività cognitive.
Nel 1936 Alan Turing con "On Computable Numbers, With An Application To The Entscheidungsproblem" definì i concetti che tuttora rappresentano i cardini degli attuali calcolatori.
Nel 1943 McCulloch e Pitts crearono il primo lavoro che può essere ricondotto all'Intelligenza Artificiale, ovvero un sistema che utilizzava neuroni artificiali con uno stato "on" e "off", ed un passaggio a "on" grazie a un determinato numero di stimoli generato dai neuroni circostanti.
Sette anni più tardi Marvin Misky e Dean Edmons costruirono la prima rete neurale di nome Snarc.
Nel 1950 Turing propone un test per vedere se un computer può essere intelligente (Test di Turing) attraverso uno specifico esperimento.
Solo nel 1956 John McCarthy coniò la parola Intelligenza Artificiale.

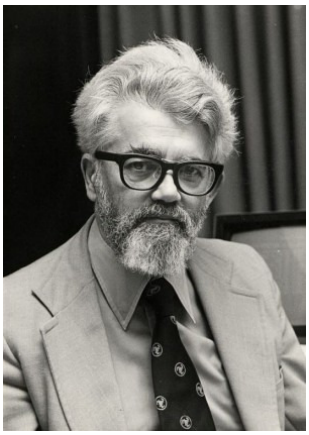

Figura 1.1: John McCarthy

Successivamente la storia dell'IA è stata caratterizzata da alti e bassi, in particolare sono stati fatti passi significativi relativamente ai modelli matematici (sempre più sofisticati per imitare alcune funzionalità cerebrali come ad esempio il riconoscimento dei pattern), ma altalenanti dal punto di vista dell'hardware e delle reti neurali.
Nel 1958 Frank Rosenblatt propone il primo modello di rete neurale ovvero il cosiddetto "Percettrone di Rosenblatt". La prima svolta importante dal punto di vista tecnologico arriva tra la fine degli anni '70 e il decennio degli anni '80 con lo sviluppo delle Gpu (graphics processing unit) che hanno ridotto notevolmente i tempi di addestramento delle reti, abbassandoli di 10/20 volte.

1.2 Cos'è l'Intelligenza Artificiale

L'Intelligenza Artificiale è un campo della scienza informatica che si occupa di realizzare macchine che possano risolvere dei problemi autonomamente, senza l'aiuto dell'uomo. È dimostrata quando un'attività, prima effettuata da un essere umano che richiede l'abilità di apprendere e ragionare, può essere eseguita in autonomia da una macchina o più in generale da un sistema intelligente.

Partendo dal funzionamento del cervello umano, di cui tuttora non riusciamo a comprenderne appieno i meccanismi, un sistema di Intelligenza Artificiale deve garantire alcune azioni/funzioni:

- *agire umanamente* (cioè in modo indistinto rispetto ad un essere umano)
- *pensare umanamente* (risolvendo un problema con funzioni cognitive)
- *pensare razionalmente* (sfruttando cioè la logica come fa un essere umano)
- *agire razionalmente* (avviando un processo per ottenere il miglior risultato atteso in base alle informazioni a disposizione)

Esistono due diverse correnti filosofiche che negli anni hanno creato una sorta di dibattito, per definire il concetto e il funzionamento di una macchina Intelligente:

- **Intelligenza artificiale debole**: sistemi tecnologici in grado di simulare le capacità umane, senza però raggiungere completamente le potenzialità intellettuali tipiche dell'uomo, ad esempio l'assistente vocale Siri
- **Intelligenza artificiale forte**: si parla di "sistemi sapienti" che possono sviluppare una propria intelligenza, non simulando quella dell'uomo ma sviluppandone una in modo autonomo; tali sistemi vengono definiti "coscienti di loro stessi"

Questo dibattito viene portato ad un livello più concreto quando nel 1950 Alan Touring pubblica un articolo chiamato "Computing Machinery and Intelligence", all'interno del quale dava la definizione di Intelligenza di una macchina basandosi su un test, chiamato "The Imitation Game" o anche "Test di Touring".

Il Test consiste in un gioco che coinvolge tre persone: un uomo (A), una donna (B) e un terzo individuo (C). C è tenuto separato e deve indovinare, facendo una serie di domande, il sesso di A e B; dall'altra parte A dovrà cercare di ingannare C, invece B dovrà aiutarlo. Il presupposto del test è che la macchina sostituisca A; se il numero di risposte corrette è uguale anche dopo la sostituzione, allora la macchina potrà considerarsi intelligente dal momento in cui sarebbe indistinguibile dall'essere umano.

La conclusione a cui arriva Touring è la definizione di macchina intelligente, ovvero di una macchina in grado di pensare, capace di elaborare idee, di metterle in relazione con altre e di saperle esprimere.

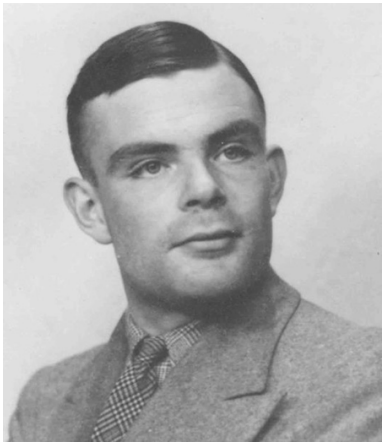

Figura 1.2: Alan Touring

Il funzionamento di un sistema di IA, dal punto di vista intellettuale, ha quattro diversi livelli funzionali:

1. **Comprensione**: attraverso la simulazione di capacità cognitive è in grado di riconoscere immagini, video, testi ed estrapolare informazioni

2. **Ragionamento**: grazie alla logica, i sistemi riescono a collegare le molteplici informazioni raccolte, utilizzando specifici algoritmi

3. **Apprendimento**: mediante funzionalità specifiche per la lettura degli input, i sistemi riescono a dare un output corretto (es. Machine Learning)

4. **Interazione**: modalità di funzionamento dell'IA in relazione alla sua interazione con l'uomo, con una forte crescita dei sistemi di NLP, Natural Language Processing, tecnologie che consentono all'uomo di colloquiare con le macchine utilizzando il linguaggio naturale

L'intelligenza artificiale è ormai parte integrante della nostra società e ha trovato una sua diffusione in diversi settori industriali, utilizzando a vari livelli le funzionalità espresse precedentemente:

- **Finanza**: da anni sono stati introdotti algoritmi di trading capaci di prendere decisioni autonome, grazie alla capacità di elaborare grandi quantità di dati in tempi molto minori rispetto ad un essere umano. Inoltre, applicazioni come ad esempio l'app Digit, che aiutano e supportano il consumatore a ottimizzare le proprie spese in base alle proprie abitudini e obiettivi personali

- **Automotive e Industria**: da decine di anni i robot sostituiscono gli esseri umani nei lavori ripetitivi e pericolosi, ma la vera evoluzione dell'Intelligenza Artificiale in questo settore è la costruzione di macchine a guida autonoma con sistemi capaci di rilevare, individuare e processare eventi e informazioni in tempo reale che consentono la sostituzione della guida umana

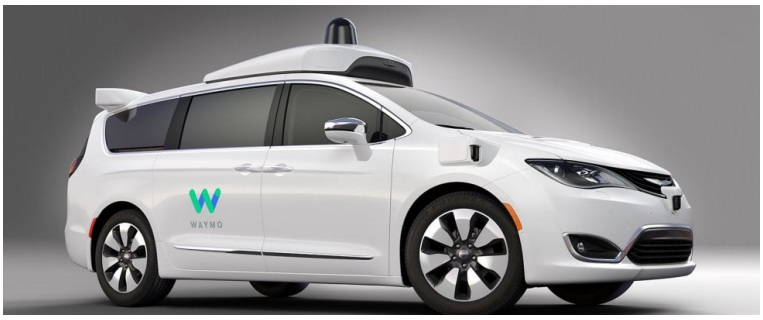

Figura 1.3: Macchina a guida autonoma

- **Vita quotidiana**: Applicazioni come Siri, Cortana, Alexa... assistono ogni giorno ciascun individuo nella ricerca di informazioni attraverso un'interazione intelligente con l'essere umano. Anche nell'ambito videoludico gli algoritmi di IA sono molto utilizzati per massimizzare l'effetto realistico del gioco e del comportamento "umano" dei diversi personaggi

- **Difesa**: negli ultimi anni sono stati introdotti robot e droni capaci di intraprendere attività autonome in sostituzione del soldato in azioni militari

- **Pubblicità**: i messaggi pubblicitari sono sempre più mirati al singolo individuo grazie ad algoritmi specifici capaci di elaborare in tempo reale le abitudini e preferenze del consumatore

- **Medicina**: questo settore rappresenta uno degli ambiti di maggior interesse e diffusione dell'Intelligenza Artificiale a supporto dell'attività del personale clinico e verrà dettagliato nei successivi capitoli

Correlati con la disciplina dell'Intelligenza Artificiale rientrano anche due ambiti di studio: il Machine Learning e il Deep Learning.

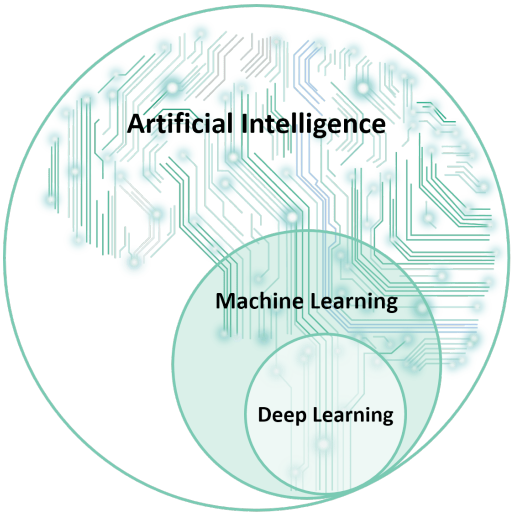

Figura 1.4: Visione d'insieme di IA,ML e DL

1.3 Machine Learning

Il Machine Learning, o apprendimento automatico, consente ai computer di "imparare" ed è di fatto un modo per raggiungere l'Intelligenza Artificiale. Il termine fu coniato da Arthur Samuel nel 1959, come "*the ability to learn without being explicitly programmed*" cioè l'abilità di un sistema di imparare senza essere espressamente programmato.
L'elemento centrale del ML è l'algoritmo, che può essere definito come un "procedimento, implementato attraverso una sequenza ordinata e finita di passi elementari, che conduce a un ben determinato risultato in un tempo finito". La base del Machine Learning è il sistema che allena l'algoritmo in modo tale da imparare come eseguire l'attività, fornendo all'algoritmo un gran numero di dati che gli consentono di correggersi o adattarsi in modo tale da migliorare sempre più le proprie prestazioni.

Il Machine Learning è caratterizzato da alcuni modelli di apprendimento ed è proprio in base a questi modelli che è possibile fare una classificazione degli algoritmi:

1. Apprendimento Supervisionato
2. Apprendimento Non Supervisionato
3. Apprendimento per Rinforzo

1.3.1 Apprendimento Supervisionato

L'**Apprendimento Supervisionato** è una tecnica di apprendimento automatico, il cui obiettivo è istruire il sistema in modo tale che possa risolvere i compiti in maniera autonoma sulla base di una serie di esempi assegnati precedentemente, formati da coppie di input e output desiderati. Al computer vengono forniti degli esempi nella forma di possibili input e i rispettivi output desiderati e l'obiettivo è quello di estrarre una regola generale che associ l'input all'output corretto.

Fondamentale nell'apprendimento supervisionato è il **training set** (addestramento) in cui viene presentato un esempio: viene generata automaticamente una risposta (output) in relazione all'input e se la risposta corrisponde a quella esatta allora si ha un rinforzo delle connessioni che hanno portato al risultato esatto. In caso contrario, ovvero se la risposta non è quella attesa, i pesi delle connessioni vengono modificati in modo da ottimizzare il risultato e diminuire lo scostamento dalla soluzione corretta. Successivamente si passa ad una fase di test in cui si verifica la capacità di generalizzazione dell'algoritmo (Test set). Il test set viene utilizzato per determinare l'accuratezza del modello.

L'**Algoritmo di classificazione** viene utilizzato in più metodi di apprendimento supervisionato e consiste nell'identificare a quale categoria appartiene una nuova osservazione, sulla base del training set dove le istanze appartengono già ad una categoria. Dunque gli algoritmi di classificazione, partendo da un training set, costruiscono un modello che verrà successivamente utilizzato per classificare le nuove istanze. Un algoritmo di classificazione è noto

come **classificatore**.

Esistono diversi Algoritmi di classificazione, di seguito vengono riportati i principali:

Alberi di Decisione

L'obiettivo dell'Albero di Decisione è creare un modello che preveda il valore di una variabile di destinazione in base a diverse variabili di input. L'Albero è formato da tre parti:

- Nodo interno: rappresenta una delle variabili di input
- Arco verso il nodo figlio: possibile valore per quella proprietà
- Foglia: rappresenta il valore di destinazione per la classe, dati i valori delle variabili di input, rappresentato dal nodo radice alla foglia

Successivamente l'Albero di Decisione viene utilizzato per stabilire a quale classe appartiene la nuova istanza che si vuole classificare. Al crescere della profondità e della complessità delle regole decisionali dell'albero si avrà a una maggior precisione dell'albero stesso.

L'Albero di Decisione è tra i migliori modelli di classificazione in quanto:

- Non è costoso da costruire
- È semplice da interpretare e capire
- È in grado di gestire sia dati numerici che di categoria
- Ha una buona accuratezza in molte applicazioni, in confronto ad altri metodi di classificazione

Ciò nonostante presenta alcuni svantaggi, come nei seguenti esempi:

- Nel caso chiamato Overfitting, con la creazione di un albero troppo complesso che non generalizza bene con i dati del training

- Nel caso di alberi poco robusti, dove un piccolo cambiamento nei dati può comportare una costruzione di un albero e previsioni finali diverse

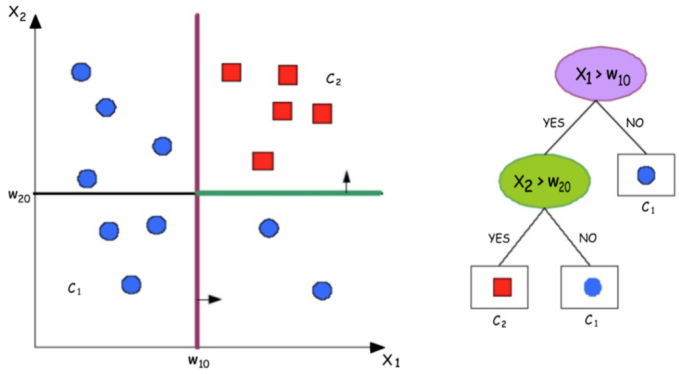

Figura 1.5: Classificazione tramite Albero di Decisione

K-nearest neighbors (k-NN)

Nel riconoscimento di pattern, il K-nearest neighbors è un metodo utilizzato per la classificazione di oggetti basandosi sulle caratteristiche degli oggetti vicini a quello considerato. È considerato il più semplice tra gli algoritmi del ML.

Fondamentale è la scelta del parametro k, il quale deve essere un numero intero, positivo, preferibilmente piccolo e dispari, per evitare casi di parità. Se k è troppo piccolo, l'approccio è sensibile al rumore, d'altra parte se k è troppo grande la classificazione può essere computazionalmente costosa e l'intorno può includere campioni appartenenti ad altre classi.
Il numero K indica gli oggetti presi in considerazione.

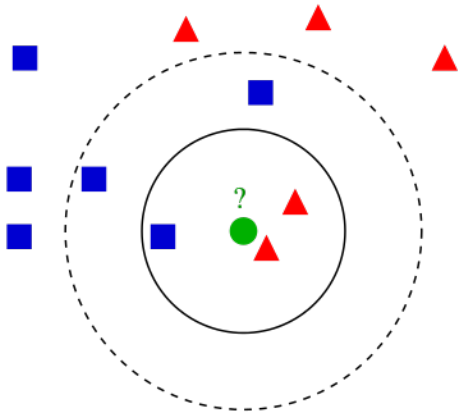

Figura 1.6: Esempio di classificazione tramite K-NN

La figura 1.6 mostra un esempio di classificazione mediante K-nn. L'oggetto sotto osservazione è il pallino verde. Abbiamo due classi:

1. I triangoli rossi
2. I quadrati blu

Nel primo caso k=3, il nostro oggetto ricadrà nella classe dei triangoli rossi, poichè ci sono 2 triangoli rossi e un 1 quadrato blu. Nel secondo caso k=5, sarà inserito nella classe dei quadrati blu, siccome sono 3 rispetto ai triangoli rossi che sono 2.

Quando i dati di addestramento sono elevati il calcolo delle distanze può risultare costoso, per evitare ciò si cerca di ridurre il numero di distanze da calcolare per la decisione.

Per il calcolo delle distanze fra gli oggetti, solitamente, viene utilizzata la distanza Euclidea, altrimenti la distanza di Manhattan, nel caso in cui a differenza dei numeri ci siano delle stringhe viene utilizzata la distanza di Hamming.

L'uso di questo algoritmo presenta diversa vantaggi:

- Il principale è che non richiede ne l'apprendimento ne la costruzione del modello
- Rispetto a sistemi basati su regole o alberi di decisione, permette di costruire "contorni" delle classi non lineari e quindi risulta più flessibile

Presenta però anche degli svantaggi, tra cui:

- la classe è determinata localmente e quindi risulta suscettibile al rumore dei dati
- sensibilità all'irrilevanza dei dati che falseranno la distanza tra gli oggetti
- il costo computazionale può essere alto per la classificazione di nuovi dati

Support Vector Machines (SVM)

Le Support Vector Machines sono metodi di classificazione che generano un'approssimazione globale del modello di classificazione utilizzando i dati di addestramento.

Consideriamo un esempio con solo due classi C =cerchi rossi, cerchi blu. Supponiamo di avere un insieme di addestramento di N campioni ognuno dei quali appartiene ad una delle due classi di C. Essendo due le classi, siamo di fronte ad un problema di classificazione binaria. L'approccio geometrico da utilizzare nel problema della classificazione binaria consiste nel trovare una superficie che separi lo spazio di input in due parti distinte, dove risiedono gli elementi appartenenti alle due classi. Nel caso i dati siano linearmente separabili possiamo effettuare una classificazione lineare in cui si assume che esista un iperpiano (ma anche più di uno) in grado di separare i campioni dell'insieme di addestramento, nell'esempio i cerchi rossi dai cerchi blu. L'obbiettivo è trovare un iperpiano che separa nel modo migliore i campioni, cioè quello che rende massimo il margine, considerato come la distanza minima

fra le due classi.

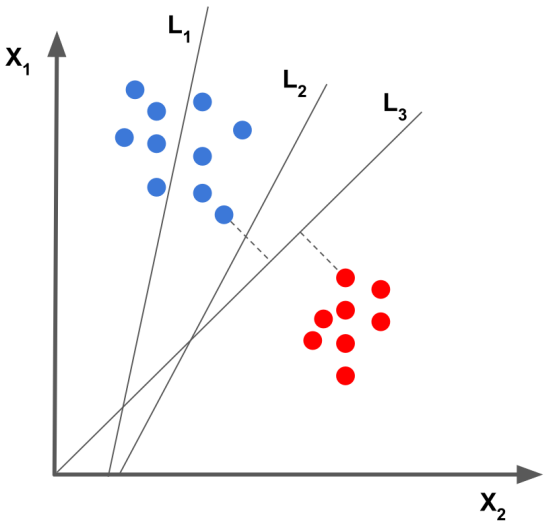

Figura 1.7: Ricerca dell'iperpiano migliore

Nell'esempio (figura1.7) abbiamo tre linee L1, L2 e L3.

- L1, non è una buona scelta perché non separa le due classi
- L2 e L3, entrambe separano le due classi, ma L3 è l'iperpiano migliore e perciò è detto Optimal separting hyperplane (OSH)

Nel caso in cui non esista nessun iperpiano in grado di separare i campioni delle diverse classi si parla di classificazione non lineare. La soluzione consiste nel proiettare l'insieme di addestramento in uno spazio di dimensione maggiore, in cui sia possibile effettuare una classificazione lineare e la ricerca dell'OSH. La funzione per effettuare questa operazione è chiamata Mapping.

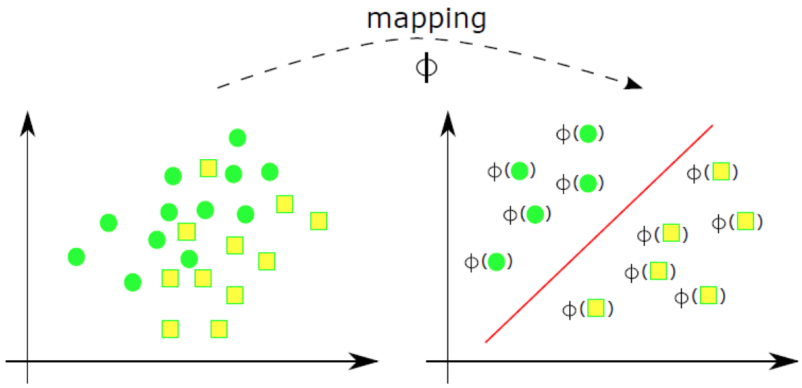

Figura 1.8: Funzione di mapping

1.3.2 Apprendimento Non Supervisionato

L'**Apprendimento non Supervisionato** è una tecnica in cui vengono forniti al sistema una serie di input che dovranno essere riclassificati in base a caratteristiche comuni per poi effettuare ragionamenti sugli input successivi. Diversamente dall'apprendimento supervisionato, i dati non sono classificati, ma successivamente devono essere scoperte automaticamente le classi. Le tecniche di apprendimento non supervisionato sono molto valide con elementi di tipo numerico, ma meno efficienti con dati di tipo non numerico.

Alcuni algoritmi di apprendimento non supervisionato sono:

Clustering

Ricerca di gruppi di oggetti tali che gli oggetti appartenenti a un gruppo siano "simili" tra loro e differenti dagli oggetti negli altri gruppi. Nell'apprendimento supervisionato questo veniva fatto dalla classificazione.
L'obbiettivo del clustering è trovare il maggiore numero di gruppi con caratteristiche omogenee, ma che siano il più possibile diversi tra di loro. Nel dettaglio, massimizzare la varianza tra i cluster, ma minimizzarla all'interno.

In biologia può essere utilizzato per derivare tassonomie di animali e piante. Nel marketing può essere impiegato per derivare e caratterizzare gruppi di consumatori.
Le tecniche di clustering possono seguire due filosofie:

1. Dal basso verso l'alto (metodi aggregativi o bottom-up): inizialmente tutti i cluster sono considerati a se stanti, successivamente l'algoritmo unisce i cluster più vicini. Il termine d'arresto è finche non si ottiene un numero prefissato di cluster, oppure fino a che la distanza minima tra i cluster non supera un certo valore o un valore prefissato.

2. Dall'alto verso il basso (metodi divisivi o top-down): al contrario di quello bottom-up, inizialmente tutti gli eleggenti sono in un unico cluster e poi l'algoritmo li suddivide in vari cluster. L'obiettivo è quello di ottenere gruppi sempre più omogenei. Il metodo si ferma quando viene raggiunto un numero prefissato di cluster.

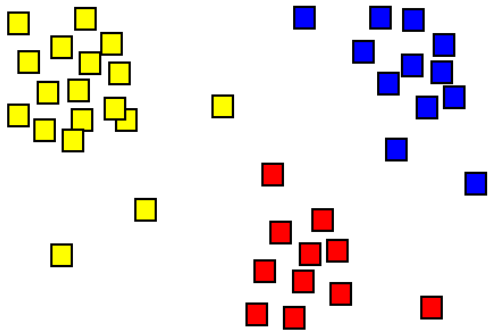

Figura 1.9: Risultato di un'analisi di cluster

La figura 1.9 mostra il risultato di un'analisi di cluster rappresentata con la colorazione dei quadrati in tre colori: giallo, rosso e blu.

Regole di associazione

È un metodo di apprendimento non supervisionato il cui obiettivo è scoprire relazioni interessanti tra variabili in database di grandi dimensioni.
Questo metodo viene utilizzato su grandi volumi di dati in cui esista il concetto di "transazione".
Le regole di associazione sono una sorta di "implicazioni".
La regola X $\Longrightarrow$ Y viene interpretata come "nelle transazioni in cui compare X compare anche Y".
X è detto corpo, Y è detta testa.

Nelle regole di associazione abbiamo due misure statistiche:

1. Supporto: che indica la percentuale di transazioni che contengono sia X che Y

2. Confidenza: date le transazioni che contengono X, indica qual'è la percentuale che contengono Y

Un esempio potrebbe essere lo scontrino dei supermercati.

Esempio:

- pane, burro $\Longrightarrow$ latte

- Supporto 0.05

- Confidenza 0.9

Afferma che:
Pane, burro e latte compaiono insieme nel 5% degli scontrini.
Gli scontrini che contengono pane e burro insieme, al 90%, conterranno anche il latte.

1.3.3 Apprendimento per Rinforzo

L'**Apprendimento con rinforzo** è una tecnica di ML che punta a realizzare algoritmi in cui i suoi agenti software siano in grado di intraprendere azioni in un ambiente in modo da massimizzare la loro ricompensa. Il problema fondamentale rimane: quali azioni devo eseguire per massimizzare la mia ricompensa futura?
La risposta non è semplice perché esistono azioni che sono utili nell'immediato, ma controproducenti nel futuro, d'altra parte azioni che sembrano inutili oggi potrebbero essere molto utili nel futuro.
Inoltre l'ambiente è, solitamente, stocastico per cui non si può avere la certezza che l'azione che andremo ad eseguire sarà produttiva.
Un dilemma importante in questo apprendimento è quello di "exploration vs explotation"; nel senso che ci si chiede quanto bisogna esplorare l'ambiente alla ricerca di nuove strategie (potenzialmente migliori) invece che concentrarsi su quelle apprese.

In molte situazioni reali ad ogni nostra azione ci viene fornito dall'ambiente un feedback, sulla bontà della nostra azione.
L'apprendimento in questo caso significa generare molte risposte casuali e osservare quali hanno il grado di bontà maggiore.

L'apprendimento con rinforzo viene utilizzato in due classi di ambienti:

- Ambienti statici (o stocastici) in cui per ogni coppia di input-output vi è un unico valore di rinforzo

- Ambienti dinamici in cui la sequenza temporale influenza l'interazione tra agente e ambiente, e modifica il valore del rinforzo e altera la frequenza con cui esso è reso disponibile all'agente

Due algoritmi tipici dell'Apprendimento con rinforzo sono:

- Temporal differences learning: è un algoritmo che si modifica al variare del tempo con lo scopo di predire le ricompense future a partire dalle conoscenze attuali

- Q-learning: è un metodo complementare che non richiede un modello dell'ambiente per predire il guadagno legato ad ogni azione; l'ambiente viene esplorato in modo stocastico e la ricompensa si ha in presenza della scelta dell'azione con il rinforzo immediato e l'utilità maggiore

1.4 Deep Learning

Il Deep Learning, in italiano Apprendimento Profondo, è un'area del Machine Learning che si ispira alla struttura ed al funzionamento del cervello biologico, ovvero della mente umana.
Possiamo considerare quindi il ML come il modello che "allena" l'IA, e il Deep Learning l'algoritmo che emula la mente umana.
L'Apprendimento Profondo, oltre agli algoritmi, necessita di reti neurali artificiali; proprio queste vogliono imitare le reti neurali del cervello umano, che sono costituite da reti di neuroni che si influenzano mediante connessioni sinaptiche.
L'IA cerca di imitare questi principi per sviluppare le reti artificiali.
Le reti neurali artificiali sono state utilizzate in vari ambiti come la Computer Vision, la traduzione automatica, il riconoscimento vocale, diagnosi mediche e videogiochi.

Ogni neurone rileva delle condizioni e le comunica attraverso la frequenza di scarica. Un neurone artificiale cerca quindi di emulare le caratteristiche di un neurone, che possono essere sintetizzate nel seguente elenco:

- I neuroni sono fortemente connessi tra loro

- Ogni connessione ha un peso, che indica la forza della connessione, e può avere un valore sia positivo che negativo

- L'input che un neurone riceve è dato dal prodotto tra il segnale proveniente da quel neurone e il peso della connessione
- L'input totale del neurone è la somma delle attivazioni che il neurone riceve dai neuroni vicini.
- L'output del neurone viene inviato a quelli "finali"

L'architettura di una rete neurale artificiale è strutturata in un determinato modo:

- Unità che direttamente dall'ambiente ricevono gli input, chiamato strato di input
- Quelle che producono l'output finale, dette strato di output
- Le unità intermedie, chiamate unità nascoste

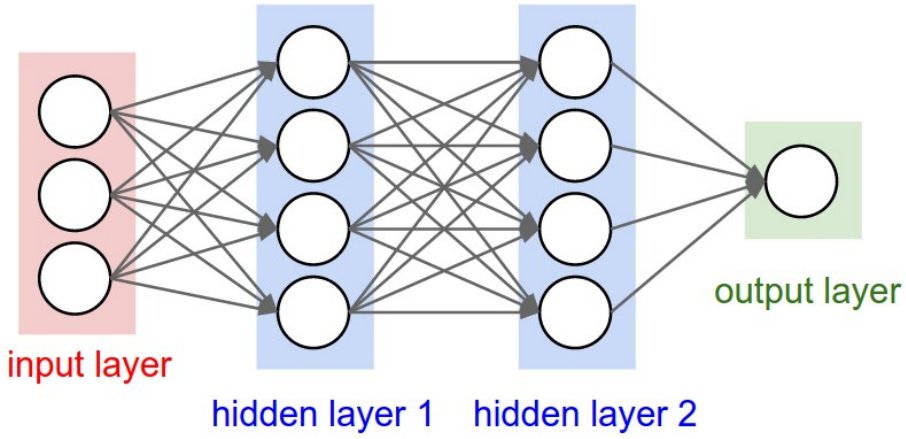

Figura 1.10: Struttura di una rete neurale artificiale

Il modo in cui i neuroni sono collegati tra loro possono essere diversi:

- Reti feed-foward con connessioni unidirezionali solo da input ad output
- Reti ricorrenti con connessioni bidirezionali

- Reti interamente ricorrenti con anche connessioni intrastrato laterali

Durante l'inizializzazione di una rete neurale il valore dei pesi sinaptici viene assegnato in maniera casuale, per poi essere modificato man mano che la macchina apprende.
L'obbiettivo principale, cioè l'apprendimento, è quello di modificare, azione dopo azione, il valore dei pesi, per far si che la predizione successiva sia più accurata di quella precedente.

Gli algoritmi di apprendimento delle reti neurali sono:

- Regola di Hebb: "se due neuroni collegati tra loro sono contemporaneamente attivi, l'efficacia sinaptica della connessione viene rinforzata". Se associamo un input x e un output y, la formula si può esprimere come: Δ w= n(costante) * y *x. Varianti della regola di Hebb, ma che comunque si basano tutte sullo stesso principio di rinforzo in base alla coordinazione della scarica:

  1. Regola postsinaptica diminuisce l'efficacia sinaptica (il valore del peso) quando il neurone postsinaptico è attivo, ma quello presinaptico no
  2. Regola presinaptica diminuisce l'efficacia sinaptica quando il neurone presinaptico è attivo, ma quello postsinaptico no
  3. Regola della covarianza è uguale a quella di Hebb, ma funziona con valori bipolari (-1/1) e non binari (0/1); ciò porta a rinforzare la connessione quando l'attività dei due neuroni è correlata mentre la indebolisce quando non è correlata

- L'idea di correggere i pesi della rete in base all'errore, calcolato come la differenza tra risultato atteso e risultato ottenuto, fu elaborata da Frank Rosenblatt. I percettroni di Rosenblatt erano delle reti neurali a due strati di connessioni, in cui la prima portava le informazioni allo strato di input, la seconda era sottoposta ad apprendimento

Per percettroni indichiamo una rete con un unico strato unidirezionale dai nodi di input a quelli di output. In un percettrone l'output si attiva quando la sommatoria pesata del prodotto tra w e x è maggiore di zero, altrimenti non si attiva. Si ha poi il confronto tra il valore ottenuto dall'attivazione e l'output atteso; se questi sono differenti si ha una modifica del peso sinaptico in base alla risposta e al tasso di apprendimento

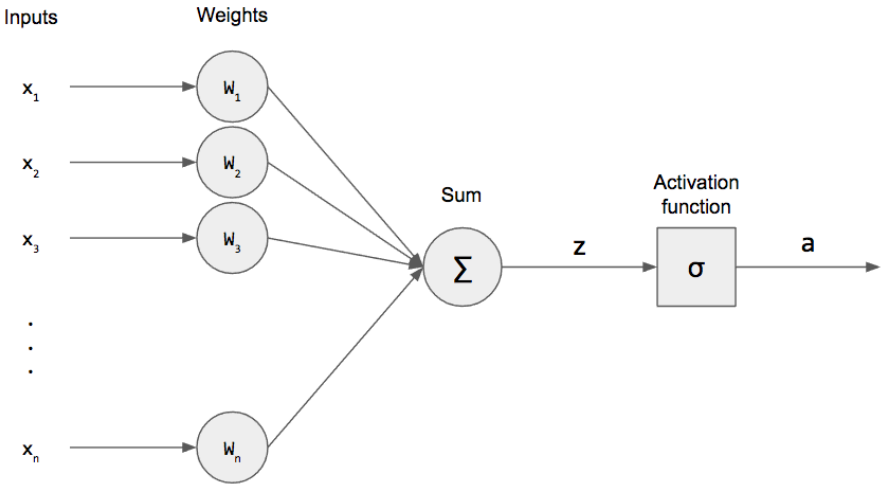

Figura 1.11: Raffigurazione di un percettrone

- La regola delta è una regola che si basa sulla modifica delle connessioni in modo da ridurre fino a zero il valore dell'errore determinato dalla differenza tra risultato atteso e risultato ottenuto.
  Questa regola permette di descrivere la prestazione con una funzione che misura l'errore della rete che si basa sullo scarto quadratico medio tra risposta desiderata e output effettivo; l'apprendimento consiste nel diminuire il valore dell'errore variando il valore delle connessioni sinaptiche

- La funzione della discesa del gradiente consente di determinare i minimi e i massimi in una funzione a più variabili. In generale, il gradiente di una funzione f, denotato con Δf,è definito in ciascun punto dalla

seguente relazione: per un qualunque vettore $\vec{v}$, il prodotto scalare $\vec{v} \cdot \Delta f$ dà il valore della derivata direzionale di f rispetto a $\vec{v}$. La ricerca basata sulla discesa del gradiente determina i vettori peso che minimizzano E (errore della rete). Ad ogni passo il vettore dei pesi è modificato nella direzione che produce la più ripida discesa del gradiente sulla superficie dell'errore

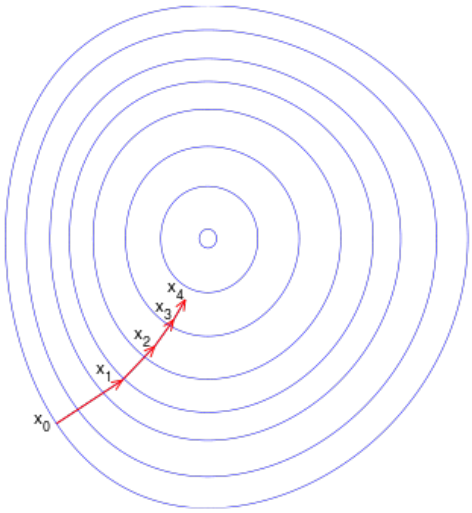

Figura 1.12: Illustrazione grafica del metodo per trovare il minimo su una superficie

- La back-propagation, in italiano retropropagazione dell'errore, è un'estensione della regola delta che permette di addestrare reti multistrato. L'algoritmo permette di modificare i pesi delle connessioni in modo tale che si minimizzi una certa funzione errore E. L'algoritmo confronta il valore in uscita del sistema con il valore desiderato (obiettivo). Sulla base della differenza così calcolata (errore), l'algoritmo modifica i pesi sinaptici della rete neurale, facendo convergere progressivamente il set dei valori di uscita verso quelli desiderati.

Capitolo 2

Applicazione del Machine Learning in Medicina

Negli ultimi dieci anni si è assistito ad un rilevante e crescente utilizzo in Medicina dell'Intelligenza Artificiale, che sta riscuotendo un grande interesse anche grazie a pubblicazioni che ne hanno rilevato una particolare precisione in specifici contesti clinici. In particolare, grazie al Deep Learning, alcuni sistemi informatici sono riusciti ad avere un'accuratezza diagnostica paragonabile a quella di medici esperti in diverse discipline. Tuttavia ad oggi mancano ancora studi ed applicazioni a largo spettro sull'efficacia riguardo obbiettivi clinici importanti come la riduzione della mortalità o il miglioramento della vita del paziente.

Numerosi sono i vantaggi che si prospettano derivare dall'applicazione del ML alla Medicina: dall'aumento della produttività del sistema, alla maggiore accuratezza diagnostica e predizione di eventi epidemici, fino alla possibilità di favorire l'accesso agli esami diagnostici anche in luoghi e a persone che non possono beneficiarne a causa di barriere geografiche, politiche ed economiche.

Ad esempio alcune applicazione dell'apprendimento automatico in sanità che portano a un beneficio dell'accuratezza diagnostica sono:

- Individuazione di tumori tramite l'analisi di immagini radiologiche
- Rilevazione della retinopatia diabetica
- Algoritmi in grado di predire eventi cardiovascolari futuri

Inoltre, anche se non ha una rilevanza ai fini clinici, l'introduzione di chatbot e assistenti vocali nel rapporto con il pubblico potrebbe migliorare il supporto ai pazienti.
Solitamente l'approccio di ML che viene utilizzato nella medicina è quello supervisionato; l'input è l'immagine diagnostica che viene fornita o la descrizione di un caso clinico e l'output è la predizione corrispondente.
L'output è tipicamente espresso come una categoria, un punteggio o un valore percentuale.
Nella fase di addestramento viene presentato al sistema un elevato numero di casi, già classificati in base al "gold standard", ovvero una diagnosi definita a maggioranza da un gruppo di specialisti.
Successivamente nella fase di test, vengono mostrate una serie di immagini, sempre classificate da esperti, ma senza che il modello sia al corrente della classificazione. In questa fase viene osservata la sua capacità predittiva e la sua accuratezza diagnostica rispetto ai casi già classificati esattamente.
Si ripete questa fase finché non vengono raggiunti livelli di accuratezza soddisfacenti.

2.1 Aspetti negativi dell'IA in Medicina

Nonostante l'utilizzo dell'IA/ML in Medicina abbia già portato numerosi risultati positivi e, come espresso nel paragrafo precedente, ci siano grandi aspettative sul suo utilizzo futuro, esistono comunque tuttora perplessità relativamente a queste tecnologie innovative a supporto dell'attività clinica-sanitaria.
Di seguito vengono elencati i principali aspetti negativi riguardo l'utilizzo dell'IA/ML in Medicina:

- Una delle principali criticità è quella dell' "over-reliance", ovvero che con il tempo i medici sviluppino un eccessivo affidamento nelle capacità dell'automazione a scapito della propria esperienza. La fiducia sarebbe alimentata dalla percezione che ogni nuova tecnologia sia migliore di quella precedente, senza una reale valutazione della sua efficacia. La conseguenza di questo eccessivo affidamento al supporto decisionale offerto dall'IA/ML potrebbe portare ad una dipendenza nell'uso di questi sistemi (overdependence), che col passare del tempo porterebbe alla dequalificazione (deskilling) del personale medico (questo aspetto verrà approfondito nel capitolo 3).

- Durante la fase di addestramento degli algoritmi vengono sottoposte al sistema immagini scelte dai medici, che sono già state considerate come corrette, cioè precedentemente classificate da parte di specialisti. Questo approccio presenta due potenziali problemi:

  1. L'eventuale differenza tra la qualità delle immagini utilizzate nella fase di addestramento rispetto a quelle presenti nelle cartelle cliniche, in base alle quali i sistemi di ML dovranno produrre le loro predizioni. La qualità delle immagini utilizzate nella fase di addestramento è alta, poichè hanno subito un processo di ripulitura e pre-elaborazione che sarebbe impossibile sostenere nella pratica clinica quotidiana. La qualità dei cosiddetti "real-word data" soddisfa poche volte i presupposti di "dato ideale" che dovrebbe alimentare un algoritmo di apprendimento automatico

  2. La possibilità che esistano più di un "gold standard" (un test campione utilizzato come riferimento per valutare altri test) nello stesso caso clinico, e ognuno porterebbe a predizioni diverse

- La difficoltà nell'attribuire le responsabilità medico-legali nel caso in cui il medico decidesse di avvalersi del supporto di sistemi di ML, che potrebbero condurlo a errori; oppure l'attribuzione della responsabilità

nel caso in cui il medico decidesse di non applicare le raccomandazioni degli stessi sistemi

- Nella storia della medicina, anche recente, l'introduzione di nuovi strumenti diagnostici o terapeutici non sufficientemente supportata da certezze scientifiche ha portato, in più occasioni, al ritiro di questi strumenti, a causa della loro inefficacia, o persino dannosità, evidenziate dall'esecuzione purtroppo tardiva di studi clinici ben condotti, secondo quel fenomeno che è stato definito "Medical Reversal"

- Un'ulteriore criticità è rappresentata dal concetto di incertezza. Un esempio può essere quello citato da uno studio recentemente pubblicato da Dharmarajan et al. focalizzato su pazienti anziani ospedalizzati per patologie cardiopolmonari acute a cui, al momento dell'accesso in ospedale, è stata effettuata una diagnosi di scompenso cardiaco, di broncopneumopatia cronica ostruttiva oppure di polmonite. Lo studio ha osservato che durante la degenza ospedaliera i pazienti ricevevano regolarmente trattamenti medici per due o più delle suddette condizioni morbose in contemporanea, e dunque non solo per la diagnosi principale effettuata al momento dell'ammissione. Questo studio esemplifica come nella pratica clinica reale i quadri clinici si collochino spesso in "aree grigie" e non siano facilmente associabili a criteri diagnostici "aurei" come riportato nei testi di medicina o nelle linee guida

2.2 Esempi di applicazione di IA/ML in Medicina

Di seguito vengono indicati a titolo esemplificativo alcuni sistemi utilizzati nelle pratiche mediche odierne:

- **DeepMind Health di Google**. È in grado di analizzare milioni di informazioni mediche in pochi minuti, velocizzando cosi i processi sanitari come l'archiviazione delle cartelle cliniche

- **Watson di IBM** è un sistema in grado di rispondere a domande espresse in un linguaggio naturale, utilizzato come sistema di supporto per le decisioni cliniche

- Uno studio della Stanford University riporta come uno specifico algoritmo riesca, sulla base di poche informazioni sulla struttura chimica del farmaco, a formulare predizioni sia sulla tossicità sia sull'instabilità della molecola, accelerando i tempi di sinterizzazione del farmaco

- Alla Stanford University è stato sviluppato un Algoritmo di Deep Learning in grado di rilevare il cancro della pelle con un'accuratezza pari o migliore a quella dei medici

Esistono inoltre delle specifiche App che supportano il paziente nella vita quotidiana:

- **Molly**: la prima infermiera virtuale al mondo sviluppata con l'obiettivo di aiutare i pazienti monitorando le loro condizioni di salute e lo stato dei trattamenti in corso

- **AirCure**: utilizza la webcam degli smartphone per confermare in modo autonomo che il paziente è in linea con le cure indicate

Capitolo 3

Casi fuorvianti del Machine Learning in Medicina

Come descritto nei precedenti capitoli, nonostante i sistemi di IA/ML si stiano progressivamente diffondendo con successo nel campo della Medicina in diverse discipline cliniche, esistono tuttora perplessità e critiche verso una adozione più spinta di queste tecnologie innovative. La potenziale barriera all'introduzione efficace e massiva dell'IA/ML a supporto delle decisioni cliniche deriva dalla natura stessa della scienza medica, che non è deterministica, ma sempre ed ancora oggi soggetta ad interpretazioni e valutazioni soggettive da parte del medico in base alla propria esperienza.
A conferma di questo contesto, di seguito vengono riportate quattro pubblicazioni scientifiche in cui vengono analizzati casi in cui i risultati conseguiti dagli algoritmi di Machine Learning si sono dimostrati fuorvianti, col rischio di supportare il medico in decisioni sbagliate.

3.1 Caso 1 - Unintended Consequences of Machine Learning in Medicine

Il primo caso preso in considerazione viene descritto nell'articolo "Unintended Consequences of Machine Learning in Medicine" di Federico Cabitza

uscito su JAMA (Journal of the American Medical Association) il 20 Luglio 2017. In questo articolo F. Cabitza riprende il concetto che già oggi le tecniche di ML hanno fatto ampi progressi in molti settori, tra cui in Medicina, e porterà ad importanti cambiamenti soprattutto nelle discipline mediche che richiedono modelli prognostici più accurati come l'oncologia e quelli basati sul riconoscimento patologico (ad es. radiologia e patologia).
Ma allo stesso tempo, all'interno del suo articolo, descrive anche quelle che definisce "conseguenze controverse" che possono derivare dall'utilizzo del ML nella pratica clinica.

In particolare l'autore si sofferma su determinati punti:

1. **Deskilling**. Una conseguenza dell'eccessiva dipendenza nell'uso di tecniche di apprendimento automatico a supporto delle decisioni cliniche nel lungo termine potrebbe essere la riduzione del livello di competenza richiesto per svolgere una funzione medica. Nel caso in cui i sistemi di ML sbagliassero o cessassero momentaneamente di funzionare, il fenomeno del deskilling risulterebbe più grave, siccome l'eccessiva dipendenza nell'uso della tecnologia avrebbe portato a una riduzione della capacità di analisi e decisionale del medico.
   Esistono già casi reali di questo fenomeno, come riportati di seguito. In un'analisi condotta da parte di un gruppo di ricercatori della City University of London sulla lettura di 180 mammogrammi da parte di 50 professionisti, è stata documentata una riduzione della sensibilità diagnostica del 14,5% per il rilievo di cancro mammario nei medici più esperti, quando a questi venivano presentate immagini di difficile lettura corredate con l'interpretazione da parte del computer, mentre solo un aumento dell'1,6% della sensibilità diagnostica è stato rilevato grazie al supporto del computer nel sottogruppo di medici meno esperti quando a questi venivano presentati casi di più semplice interpretazione. Un altro studio su 30 pazienti ha mostrato una diminuzione nell'accuratezza diagnostica dal 57% al 48% quando gli elettrocardiogrammi erano

annotati con diagnosi computerizzate non accurate.
Questi risultati evidenziano come l'eccessivo affidamento a tecniche di ML porti ad influenzare in modo significativo le performance del personale medico, ma anche che è necessaria ancora molta ricerca per individuare in modo preciso le caratteristiche di questo fenomeno, in particolare la perdita della "fiducia in se stessi" con la conseguente incapacità a fornire interpretazioni e diagnosi definitive.

2. **Decontestualizzazione dei dati**. Gli algoritmi di ML si nutrono di grandi volumi di dati, perché solo grazie a questi possono elaborare le loro predizioni a supporto delle decisioni. Affidarsi però a questi sistemi richiede che i dati alla base delle analisi siano considerati come rappresentazioni affidabili e complete dei fenomeni che dovrebbero rappresentare in una forma discreta. Questo aspetto può costituire un problema quando il contesto clinico per sua natura non è facilmente rappresentabile, al contrario dei dati che sono di facile codificazione mediante numeri. Il problema sorge quando i sistemi di ML devono gestire informazioni di contesto difficilmente "datificabili" (rappresentabili tramite dati), quali per esempio aspetti culturali, sociali o psicologici di un paziente, oppure aspetti organizzativi di un contesto ospedaliero, che sono necessari per una corretta terapia di cura, ma che non essendo disponibili ai medici potrebbero indurre a decisioni scorrette.
Un esempio in cui la capacità predittiva di un algoritmo di ML a supporto delle decisioni è risultata tecnicamente valida, ma potenzialmente fuorviante, è stato illustrato da Caruana et al. nell'ambito di una casistica di 14.199 pazienti con polmonite, su cui sono stati applicati differenti algoritmi di ML allo scopo di predire il rischio di mortalità e indirizzare così la gestione dei pazienti in ambito intra- o extra-ospedaliero. Gli algoritmi hanno analizzato che erano a minor rischio di mortalità i pazienti affetti da asma e polmonite, rispetto a quelli che avevano solo polmonite, il quale può sembrare sorprendente. I medici infatti esclusero che l'asma potesse essere un fattore protettivo per tali pazienti

attribuendo l'errore a un "intervento medico confondente". Bisogna ricordare che i modelli di apprendimento automatico non applicano regole ai dati forniti, ma cercano relazioni sottili all'interno di essi.
La causa che induceva l'algoritmo a produrre questi risultati, anche se corretti, era dovuta al fatto che negli ospedali in cui si svolgeva lo studio, i pazienti con polmonite e storia di asma erano ricoverati in terapia intensiva e presentavano una minore mortalità, grazie ad un maggior controllo clinico.
Quindi, la mancanza del contesto clinico nell'inclusione degli algoritmi di ML ha portato a supporre che pazienti affetti da polmonite ed asma avessero esiti migliori di quelli a cui era stata diagnosticata la sola polmonite, con una riduzione del tasso di mortalità del 50% circa (5,4% vs 11,3% rispettivamente).
Questo dimostra come la mancanza di una variabile nei dati, nell'esempio l'ammissione del paziente in terapia intensiva, possa indurre gli algoritmi, benchè perfetti, a sbagliare a causa della parzialità dei dati.
In generale si può concludere che la mancata inclusione di fattori clinici difficili da rappresentare nei sistemi di ML potrebbe condurre a errori contestuali, e che quindi un'elevata dipendenza da questi sistemi potrebbe generare decisioni errate con maggior frequenza e conseguenze negative

3. **Rischio del "Black Box"**. Gli algoritmi di Machine Learning vengono spesso definiti come "modelli Black Box" nel senso che l'output generato è difficilmente comprensibile non solo dai medici utilizzatori di questi sistemi, ma anche dagli ingegneri che li hanno sviluppati. Alla luce dell'elevata accuratezza predittiva degli attuali modelli di Deep Learning, associata alla loro assenza di trasparenza, e alla loro iperscrutabilità si può ipotizzare come questi dispositivi, qualora diventassero di uso comune, potrebbero influenzare in maniera rilevante numerosi aspetti della decisione medica, andando persino a generare, nel lungo periodo, una sorta di "affidamento oracolare", ovvero di eccessiva fidu-

cia e quindi di potenziale dipendenza da questi sistemi. Ciò potrebbe modificare il modo in cui i medici apprendono, pensano, agiscono e interagiscono con colleghi e pazienti. Per alleviare la tensione tra accuratezza e interpretabilità, si stanno cercando di sviluppare sistemi di ML che forniscano in modo autonomo spiegazioni ai medici offrendo strumenti interrativi per esplorare le implicazioni di potenziali variabili di esposizione

3.2 Caso 2 - Unleashing the true potential of social networks: confirming infliximab medical trials through Facebook posts

Il secondo caso preso in considerazione viene descritto nell'articolo "Unleashing the true potential of social networks: confirming infliximab medical trials through Facebook posts" di Marco Roccetti, Catia Prandi, Paola Salomoni, Gustavo Marfia.

La ricerca di informazioni mediche online sta diventato sempre più frequente al giorno d'oggi, infatti molte persone cercano su Internet sintomi, consigli o rimedi medici dopo aver avuto infortuni o problemi di salute. Questo comportamento può risultare molto rischioso, in quanto non si ha nessuna garanzia riguardo la qualità delle informazioni che possono essere trovate online, specialmente su forum non moderati e social network. L'avvento di questi tool online ha cambiato radicalmente lo scenario nei confronti delle strutture assistenziali; i pazienti sono posti di fronte a molti più stimoli rispetto a quelli a lungo forniti dai professionisti del settore sanitario o dai tradizionali gruppi di supporto.
Nonostante l'ampia accettazione che gli strumenti online hanno ricevuto dalle comunità di assistenza sanitaria e dai pazienti, questo scambio di informazioni può determinare alcuni problemi, come la presenza di consigli medici

errati, imprecisi, incompleti, impropriamente enfatizzati, ambigui o contestabili.

Tuttavia la conoscenza delle esperienze altrui può supportare e rafforzare positivamente la fiducia del paziente, confermare le terapie di trattamento, fornire nuove alternative quando ci si trova ad affrontare problematiche e contribuire ad alleviare la solitudine mantenendo relazioni con gli altri.
È quindi fondamentale valutare la qualità delle informazioni online, sia per chi cerca determinati argomenti per la prima volta, sia per approfondire la comprensione della malattia nell'ambito di specifiche comunità.
Un fenomeno interessante che si verifica è il "prosumerismo", in quanto la condivisione online di informazioni, stimolata dall'anonimato, può aumentare la franchezza e la sincerità quando vengono toccati i problemi e le esperienze personali. Questo può essere vero sia per i pazienti sporadici che per quelli cronici; i primi cercano informazioni riguardo a sintomi sconosciuti, mentre quelli cronici ricercano aggiornamenti e nuove soluzioni ai loro noti problemi. Proprio questi ultimi trascorrono buona parte della loro vita ottimizzando il trattamento e il loro stile di vita, per raggiungere una qualità della vita più alta, acquisendo un profondo "know-how" e sviluppando un alto grado di autocoscienza su come devono essere gestite le malattie rispetto al paziente occasionale.
La condivisione online delle proprie esperienze potrebbe essere utile sia per i nuovi pazienti, ma anche per i ricercatori medici: in quanto il primo potrebbe imparare a gestire nuove situazioni, mentre il secondo potrebbe acquisire una conoscenza più profonda riguardo la propria area clinica di studio.
Questo articolo si concentra sul ruolo di un popolare strumento di social networking online, come Facebook, per una particolare classe di malati cronici, la malattia di Crohn (CD). Tale malattia può essere trattata, ma non definitivamente curata; tipicamente le persone che ne soffrono combattono per tutta la loro vita con i loro sintomi, potenzialmente condividendone online informazioni ed esperienze.

È stata eseguita un'analisi focalizzata sul dibattito online della malattia di Crohn, dove sono stati studiati i commenti dei pazienti in relazione a specifici argomenti, come cause sintomatiche possibili, sintomi, trattamenti e effetti collaterali. Quest'analisi ha portato a due principali risultati: le informazioni sul CD sono più facili da trovare sulla pagina Facebook, piuttosto che su Twitter, e il trattamento farmaceutico che genera il maggior numero di post è l'infliximab.

L'analisi è stata eseguita in tre diversi steps:

1. Attraverso metodologie di data mining, analizzando il sentimento che emerge in corrispondenza della discussione di un certo argomento
2. Attraverso la comprensione di come la discussione di un certo trattamento online influenzi l'umore, che può essere rilevato da post. In questo secondo step, l'infliximab è stato individuato come il trattamento che influenza maggiormente l'umore dei pazienti
3. Attraverso la lettura e l'interpretazione uno ad uno dei post riguardati l'infliximab, cercando una o più regole che caratterizzassero il comportamento, la percezione e i commenti pubblicati online. Queste informazioni sono state messe in relazione con la letteratura scientifica più significativa relativamente all'uso dell'infliximab

Per rilevare i post dove le persone discutono di CD su Facebook, sono state individuate le pagine pubbliche dedicate a questa malattia attraverso un metodo quantitativo.
L'analisi si è quindi sviluppata con il seguente controllo di integrità: una classificazione dei 20 autori più prolifici, leggendo il contenuto dei loro post. Il risultato di questa analisi è che la maggior parte degli autori su Facebook sono pazienti, in quanto di solito condividono esperienze personali relative al loro percorso terapeutico, al contrario di quanto succede su altri social, come Twitter.

Il problema consisteva nel trovare un modo affidabile per valutare il sentimento dell'utente relativo ad un post nel quale veniva citato il farmaco. Per eseguire ciò è stato utilizzato OpinionFinder, un sistema che può elaborare un corpo di testo e frasi identificative, restituendo un valore che soddisfa il sentimento di ogni frase, classificandolo come neutro, positivo o negativo. Quindi OpionionFinder è stato utilizzato con l'obiettivo di misurare come un sentimento espresso da un paziente possa cambiare in relazione alla somministrazione di uno specifico farmaco. Infatti la rilevanza di questo lavoro non è trovare un valore assoluto del sentimento, ma la variazione dell'umore espresso dai pazienti nei vari post.
A questo punto il secondo step è stato costruito con l'obiettivo di verificare come la discussione di un certo trattamento influenzi/causi umori positivi o negativi, partendo da un'analisi sui post di Facebook.
Il risultato di tale analisi, anche utilizzando l'analisi di Granger, ha individuato che principalmente l'infliximab influenza l'umore in positivo o negativo, senza però essere capace di decidere per uno o per l'altro.
Per questo motivo si è esteso il lavoro con un'analisi qualitativa per analizzare i post attraverso un approccio induttivo; questa tecnica è stata utilizzata per analizzare i dati raccolti dai social media con lo scopo di identificare temi comuni/non comuni alla ricerca medica tradizionale. L'interessante risultato dello studio è stato notare come sia possibile trovare punti comuni e identificare norme che emergono da entrambi i contesti.

Nel dettaglio lo studio parte con un'analisi quantitativa sui trattamenti farmacologici per i malati di CD attraverso due domande:

1. Quanto positivo o negativo le persone considerano un trattamento dai loro post online?

2. In che modo l'umore espresso dai pazienti online è influenzato dall'uso di determinati farmaci?

Nell'analisi sono stati presi in considerazione tutti i farmaci con lo stesso

principio attivo.

Per visualizzare l'evoluzione sentimentale di un certo trattamento con un grafico, il valore sentiment è stato calcolato come la differenza tra i sentimenti positivi e negativi espressi durante una settimana. Per esempio se OpinionFinder trova 10 positivi e 4 negativi, il valore del sentimento per quella settimana sarà 6.

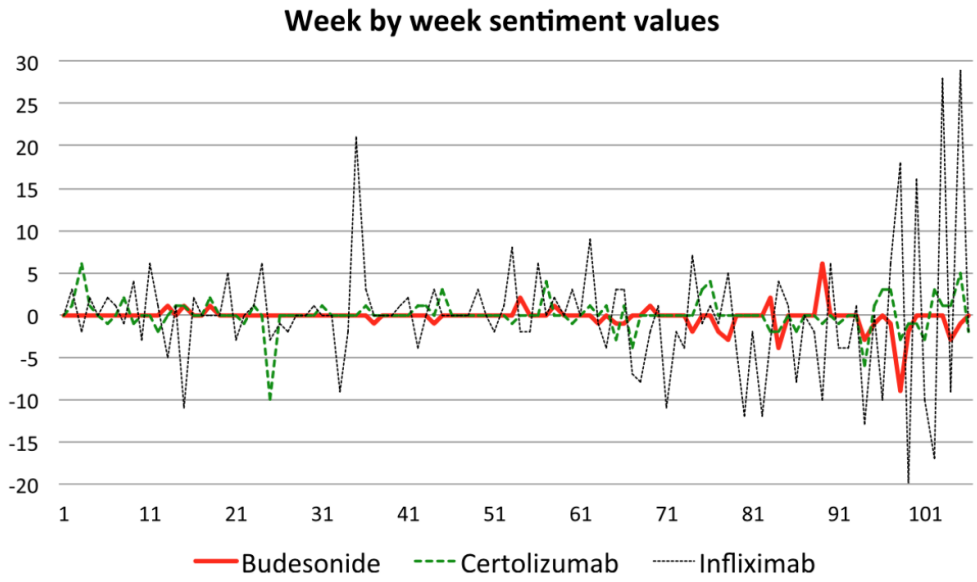

Figura 3.1: Discussioni riguardo Bedesonide, Certolizumab e Infliximab

In figura 3.1 possiamo notare come le discussioni riguardo l'infliximab risultano neutrali per un lungo periodo poi alla fine si accendono, con sentimenti positivi che si sovrappongono a quelli negativi.

Come ulteriore verifica sono state contante le parole (sentiment) positive e quelle negative per tutti i trattamenti, come mostrato in figura.

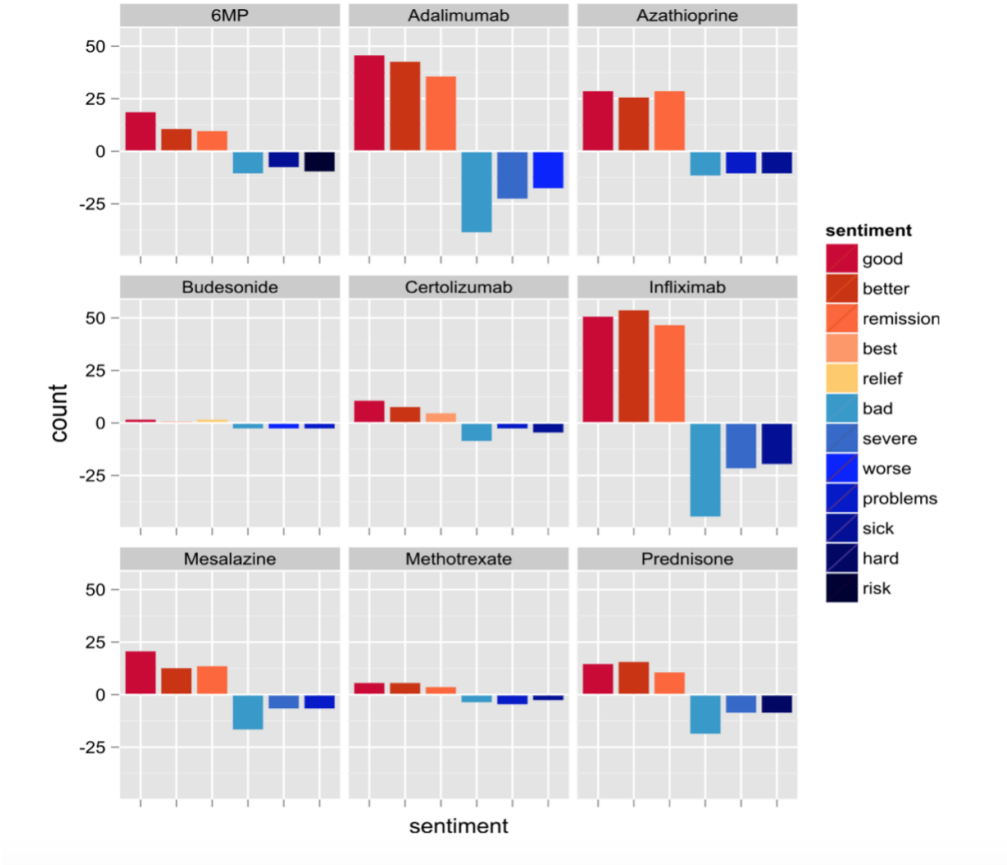

Figura 3.2: Risultati del conteggio delle parole positive e quelle negative

I risultati mostrati in figura 3.2 hanno portato alla decisione di esaminare solo tre farmaci (adalimumab, azathiorphine e infliximab) anche se quest'analisi non ha portato ad una comprensione esaustiva dell'umore espresso nei post.
Si è quindi proceduto a una valutazione delle opinioni espresse durante le discussioni di un certo trattamento ad esempio verificando se le discussioni influenzano/causano un cambiamento dello stato d'animo sulle pagine di Facebook riguardo il CD.

Nella ricerca di una relazione tra una discussione di un trattamento farmacologico e l'umore positivo/negativo si è utilizzato l'analisi di Granger e solo l'infliximab ha dimostrato una correlazione statistica significativa.
Concludendo l'infiliximab è in grado di influenzare la comunità dei malati di morbo di Crohn, ma non si è sicuri se in modo positivo o negativo.

Il risultato controverso dell'analisi quantitativa dei post di Facebook porta qui a un'ulteriore investigazione in base a come l'infliximab influenza realmente l'umore della comunità affetta da morbo di Crohn.
Sono stati analizzati i post su Facebook, in totale 241, di cui 115 sono stati scartati perché considerati neutrali. Tale analisi ha cercato, laddove possibile di classificare l'umore finale di ciascun post.
È stata utilizzata una scala Likert compresa tra [-2,2] in cui i valori rappresentavano:

- Molto insoddisfatto (-2)
- Non soddisfatto (-1)
- Incerto (0)
- Soddisfatto (1)
- Molto Soddisfatto (2)

	Very dissatisfied	Dissatisfied	Unsure	Satisfied	Very satisfied
Number of posts	6	35	22	47	16
% of posts	4.8	27.8	17.4	37.3	12.7

Figura 3.3: Risultati di soddisfazione con l'infliximab

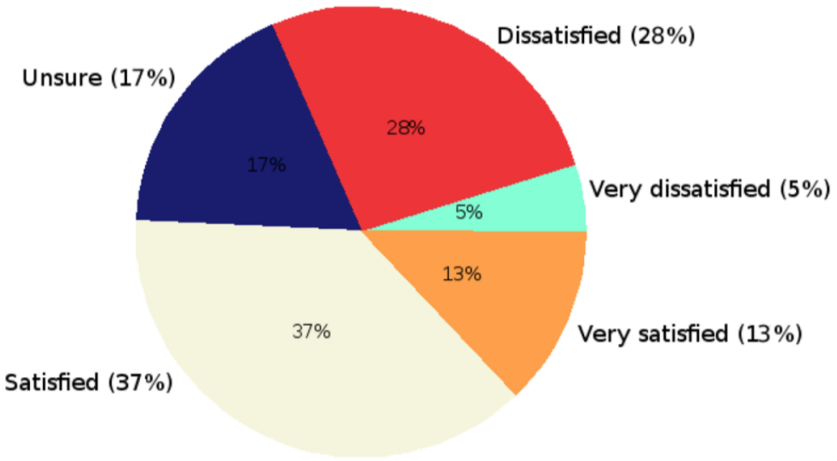

Figura 3.4: Risultati di soddisfazione con l'infliximab

Le figure 3.3 e 3.4 mostrano come non sia possibile determinare se l'utilizzo di l'infliximab provochi un sentimento positivo o negativo, in quanto il grado di soddisfazione è equamente distribuito.
Successivamente è stata condotta un'analisi tenendo conto anche della durata del trattamento; è stato possibile estrarre, oltre a un valore che rappresenta l'umore del post, anche un'altro valore numerico che rappresenta la durata del trattamento, per cercare una correlazione tra la durata del trattamento e l'umore dei pazienti.
Sorprendentemente i dati appaiono molto più interessanti raggruppati in base alle diverse durate di trattamento, rilevando due risultati:

1. Quasi il 30% dei pazienti esprime un grado di insoddisfazione rispetto all'uso di l'infliximab

2. Una durata del trattamento più lunga corrisponde ad un livello superiore di soddisfazione

Analizzando caso per caso la durata dei trattamenti si è evidenziato che:

- Nel caso di un trattamento della durata di un mese ha portato ad una insoddisfazione verso l'infliximab

- Nel caso di un trattamento della durata compreso tra 2 mesi e 2 anni sono stati contati più post positivi rispetto negativi, con un lieve incremento della tendenza di soddisfazione
- Nel caso di un trattamento della durata compreso tra 2 e 5 anni, il risultato è inaspettatamente 50% - 50%.
- Nel caso di un trattamento della durata compreso tra 5 e 13 anni, i pazienti sono chiaramente soddisfatti

È stata condotta una revisione della letteratura medica per verificare se i risultati descritti potrebbero essere significativi anche da un punto di vista medico. Un primo dato emerso è la correlazione diretta tra l'esiguo numero di pazienti coinvolti nello studio clinico e il gran numero di risultati positivi. Una possibile spiegazione potrebbe essere la maggior capacità dei medici di controllare l'evoluzione della malattia e la maggior adeguatezza per ogni singolo paziente di un certo trattamento. Al contrario quando il numero di pazienti coinvolti in uno studio cresce aumenta anche l'insoddisfazione al farmaco.

La conclusione del lavoro riporta risultati controversi relativamente a pazienti con malattia di Crohn che assumono l'infliximab; tali risultati possono essere compresi quando vengono prese in considerazione la durata del trattamento e la dimensione della popolazione coinvolta nello studio. Infatti esiste un parallelismo tra le due sorgenti di dati (post su Facebook e studi clinici descritti nella letteratura medica aggiornata) ovvero che al crescere della dimensione della popolazione analizzata cresce anche la percentuale di pazienti che riportano un'esperienza negativa.
Un possibile sviluppo di questo lavoro sarebbe quello di estendere le informazioni di base sui pazienti quali ad esempio età, sesso, altre malattie, che consentirebbero un'analisi più efficace; sfortunatamente ad oggi i social media non consentono nessun tipo di deduzione che possa essere considerata scientificamente valida.

Quindi ad oggi l'analisi di una reta di social media come Facebook non può supportare la comunità medica con nessuna informazione aggiuntiva che sarebbe viceversa molto utile alla comunità medica per facilitare i propri studi.

3.3 Caso 3 - The need to approximate the use-case in clinical Machine Learning

Il terzo caso preso in considerazione viene descritto nell'articolo "The need to approximate the use-case in clinical machine learning", che riporta quanto l'accuratezza di previsione sia fondamentale nella corretta esecuzione negli algoritmi di ML .
L'approccio normalmente utilizzato per verificare tale accuratezza è la convalida incrociata, in inglese Cross-Validation (CV).
In questo articolo vengono confrontati due popolari metodi di CV: subject-wise e record-wise. In questo metodo l'insieme dei campioni viene diviso in K partizioni della stessa dimensione. Ad ogni iterazione del metodo una parte della partizione viene utilizzata per il test e le rimanenti per il training, con cui si costruisce il modello. Lo scopo è quello di valutare la capacità dell'algoritmo di generalizzare nuovi dati.
Affinché la CV sia valida, i set di training e test devono essere indipendenti. Il concetto di indipendenza dipende dallo scenario d'uso:

- *Diagnosi*: si sviluppano modelli globali che possono essere utilizzati per nuovi soggetti, il metodo di CV utilizzato deve essere subject-wise nel senso che il training e il test set devono contenere record di diversi soggetti

- *Prognosi*: si necessita di modelli personali che possano predire futuri stati clinici di un determinato soggetto, dividendo i dati in base al tempo, in modo tale il training e il test set contengano record dello

stesso soggetto in diversi periodi, quindi il metodo utilizzato è record-wise

Pertanto la scelta del metodo di CV dipende dal caso d'uso.

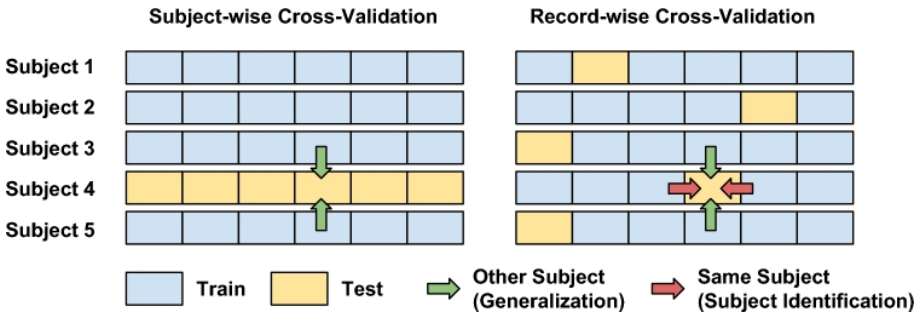

Figura 3.5: Visualizzazione di CV subject-wise e CV record-wise

La figura 3.5 mostra uno scenario di diagnosi, in cui si vuole costruire un modello che poi verrà utilizzato per classificare nuovi soggetti; quindi occorre utilizzare la CV subject-wise.

Ciononostante, si può utilizzare anche il metodo record-wise , il quale suddivide casualmente i dati in training e test set indipendentemente dai soggetti a cui appartengono. Quindi i record dello stesso soggetto si trovano sia nella fase di training che di test. In questo modo l'algoritmo può trovare un'associazione tra le caratteristiche uniche di un individuo (come ad esempio la velocità di camminata) e il suo stato clinico, che migliora automaticamente la precisione di predizione.
Il metodo record-wise CV può portare a una sovrastima dell'accuratezza di predizione dell'algoritmo.

Nell'articolo viene fatto un esempio che chiarisce perché l'utilizzo del metodo CV record-wise in una diagnosi risulta essere fuorviante.

Il caso analizzato considera quattro individui, due sani e due affetti da Par-

kinson; l'algoritmo di apprendimento automatico stima se una persona è affetta da Parkinson in base alla sua velocità di camminata.
I soggetti sani hanno rispettivamente una velocità media di 1 m/s e 0,4 m/s e quelli malati 0,6 m/s e 0,2 m/s.
Utilizzando il metodo CV subject-wise non saremo in grado di prevedere le prestazioni del soggetto sano ne di quello malato, ottenendo un'accuratezza di predizione all'incirca del 50%.
Nel caso invece in cui si utilizzi una CV record-wise , si hanno a disposizione 10 record per ogni soggetto, 9 misurazioni vengono utilizzate per predire la 10ima.
Saremo quindi in grado di sapere la condizione del soggetto.
Mediante questo metodo, l'identificazione dell'individuo sarebbe relativamente facile, e sostituirebbe il riconoscimento della malattia, che sappiamo essere invece complessa.
In questo caso il metodo CV record-wise darebbe un'accuratezza del 100%, che chiaramente non è supportata dai dati, e quindi l'algoritmo non generalizzerà.

Nell'articolo viene condotta un'altra verifica per confermare la differenza tra i due metodi presentati di convalida incrociata record-wise e subject-wise. Il dataset impiegato prende in considerazione un set di riconoscimento delle attività umane, disponibile pubblicamente, il quale contiene registrazioni di 30 soggetti nell'esecuzione di 6 attività: seduto, in piedi, camminando, salendo e scendendo le scale e sdraiato.
I dati utilizzati sono stati raccolti mediante registrazioni dall'accelerometro e dai sensori del giroscopio dello smartphone.
Nell'articolo, come classificatore, è stato utilizzata la foresta casuale, un insieme di alberi decisionali, dove ogni singolo albero fornisce una predizione sulla classe dei dati in input. Il risultato finale della foresta casuale viene determinato aggregando tutte le predizioni dei singoli alberi.
Ogni singolo albero di una foresta casuale si riferisce solo a un sottoinsieme

di caratteristiche e di campioni di dati in input. Una foresta casuale ha quindi un minor numero di parametri da ottimizzare, rendendo questo metodo meno incline all'overfitting e un ottimo metodo per classificare nuovi dati.

Per entrambi i metodi di convalida incrociata sono stati utilizzati 2, 10 o 30 soggetti e k-fold CV con k= 2, 10 o 30. Per quanto riguarda subject-wise, sono stati suddivisi i dati in modo tale che la fase di training e di test contenesse record di soggetti diversi.
In entrambi i metodi, il classificatore è stato addestrato su tutti i record fuorchè uno, su cui è stata condotta la fase di test.

Inizialmente si sono valutate le prestazioni del metodo di convalida incrociata subject-wise.
Nel momento in cui si sono utilizzati 2 fold e 2 soggetti il tasso di errore dell'algoritmo di classificazione inizialmente rilevato al 27% ma, aumentando il numero di soggetti presi in considerazione fino ad un massimo di 30, esso è diminuito raggiungendo il 9%. Aumentando il numero di fold, fino ad un massimo di 30, i dati di più soggetti hanno addestrato il classificatore e il tasso di errore ha raggiunto il 7%.
Successivamente è stato utilizzato il secondo metodo (record-wise) e usata la stessa procedura per analizzare come l'errore variasse in base al numero di fold e soggetti. È interessante osservare come l'errore di classificazione di partenza, anche nel momento in cui sono stati presi in considerazione dati di soli due soggetti e fold, era già al 2%, e aumentando il numero di questi (soggetti e fold) l'errore non è variato in modo significativo.
Da sottolineare che utilizzando questo metodo è stata sovrastimata la precisione di classificazione su questo dataset.

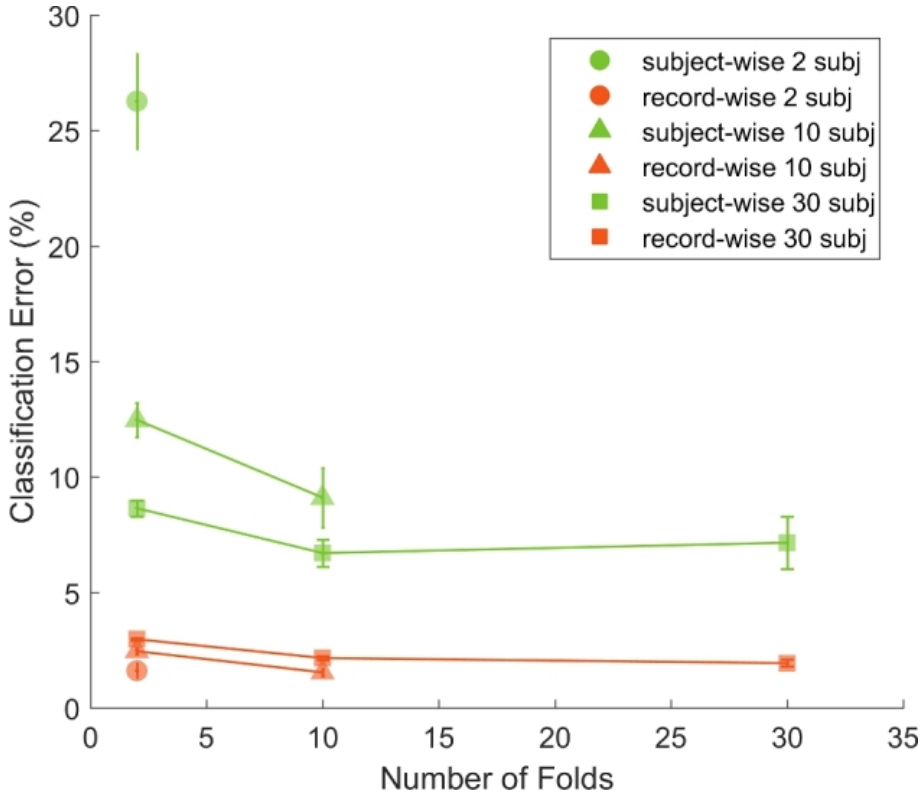

Figura 3.6: Risultati del test sulle attività umane

Il tema centrale dell'articolo è valutare l'affidabilità dell'accuratezza del ML e la tecnologia dei sensori wearable per predire risultati clinici. Prendendo in considerazione il dataset, è stato verificato come il metodo di convalida incrociata record-wise produca un'accuratezza elevata, ma fuorviante.
Nell'articolo è stato considerato solo un modo in cui la convalida incrociata possa essere usata impropriamente nell'apprendimento automatico. In molti algoritmi di predizione, oltre ai parametri principali, devono essere considerati anche degli iper-parametri, che molto spesso sono scelti in modo tale da minimizzare l'errore di predizione sul set di dati. Il modo corretto per regolare questi iper-parametri è suddividere ulteriormente il training set in sottoinsiemi di training e validation set e minimizzare l'errore di predizione in questi sottoinsiemi rispetto a farlo sull'intero set di dati. Questo approccio consente una corretta valutazione della generalizzabilità dell'algoritmo.

Mentre nell'articolo è stato utilizzato solo un dataset, è possibile utilizzare qualsiasi set di dati clinico per mostrare come la miscelazione tra soggetti nella fase di training e di validation set si possa aumentare, in modo artificiale, l'accuratezza della previsione.
Siccome in Medicina i registri clinici spesso includono informazioni sulle caratteristiche fisiche o fisiologiche di un individuo, come il girovita o il gruppo sanguino, che non variano molto nel tempo, se utilizziamo il metodo record-wise, gli algoritmi sapranno già l'esito clinico per un individuo con caratteristiche specifiche. D'altro canto l'altro metodo assicurerà che non ci siano scorciatoie per l'algoritmo.
Pertanto la scelta del metodo di convalida incrociata è particolarmente importante nelle applicazioni di previsione clinica.

Dal momento che gli algoritmi di ML sono sempre più utilizzati, è necessario valutare mediante la CV la loro accuratezza, poichè procedure di convalida non adeguate potrebbero portare a risultati inaffidabili, che condurrebbero al problema di irriproducibilità dei risultati della ricerca e quindi minare la fiducia nella medicina e nella scienza dei dati.

3.4 Caso 4 - Meaningless comparisons lead to false optimism in medical Machine Learning

Il quarto caso preso in considerazione è ripreso da un articolo pubblicato il 26/09/2017 intitolato "Meaningless comparisons lead to false optimism in medical machine learning".
Nell'articolo si evidenzia il ruolo fondamentale svolto dai big data nella diffusione di algoritmi di ML a supporto delle decisioni cliniche in Medicina. Un modo semplice ed economico per estrarre il maggior numero di informazioni mediche più precise e migliorare la personalizzazione del monitoraggio, della

diagnostica e dei trattamenti di assistenza sanitaria è l'utilizzo di dispostivi come smartphone e smartwatch.

L'articolo si concentra sul benessere mentale essendo una patologia che richiede un monitoraggio a lungo termine e che quindi viene reso sostenibile grazie ai dispositivi e sensori di uso comune.
I tipici algoritmi relativi a previsioni sul benessere mentale apparentemente sembrano funzionare bene, ma in realtà è solo perché stanno indovinando gli stati medi personali degli individui. Questo esempio sottolinea come sia facile ottenere risultati falsamente ottimistici nel momento in cui il risultato ottenuto dall'algoritmo viene confrontato con la baseline della popolazione rispetto a quella del singolo individuo.
Questo approccio viene seguito dalla maggior parte degli studi (circa il 77%) in letteratura, viene utilizzato come metodo di paragone la baseline della popolazione. L'articolo propone una nuova misura, lo "user lift", che misura il beneficio dell'algoritmo relativo al modello di una sola persona per evitare di ottenere conclusioni falsamente ottimistiche.

Per valutare l'esattezza dell'algoritmo nel predire un risultato binario, come per esempio una giornata felice rispetto ad una triste, viene considerato l'errore di predizione, che è la percentuale di osservazioni che sono state erroneamente predette. Viceversa, per valutare l'esattezza di un algoritmo per predire il livello di felicità o di stress di un individuo, viene utilizzato l'errore quadratico medio (RMSE), che indica la discrepanza quadratica media fra i valori dei dati osservati ed i valori dei dati stimati.
In entrambi i casi, errori con valori bassi indicano che l'algoritmo sta prevedendo bene lo stato di un individuo, invece valori più alti indicano un errore di predizione più significativo.

Le due baseline che prendiamo in considerazione sono quella personale, cioè di un singolo individuo, e quella della popolazione. Quella personale suppone che ogni individuo sia sempre in uno stato costante, ma può differire da persona a persona, invece quella della popolazione prevede che tutti gli individui siano nello stesso stato.

Viene proposto il metodo "user lift" per valutare se l'algoritmo sta facendo migliori predizioni rispetto a semplicemente indovinare lo stato dell'individuo. È il miglioramento delle previsioni dell'algoritmo rispetto alla baseline personale.

L'articolo presenta un esempio di come l'utilizzo di tecniche di ML possano condurre a conclusioni falsamente ottimistiche, in particolare nella predizione di stati mentali di benessere di un individuo. Sono stati considerati due dataset, entrambi con dati raccolti dagli smartphone dei soggetti analizzati, in particolare la loro posizione GPS e i loro livelli di stress e felicità:

- **Dataset StudentiLife**: relativo ad un gruppo di studenti di un'università Americana vengono raccolte le misure giornaliere di stress su una scala Likert a cinque punti
- **Dataset MIT Friends and Family**: relativo ad un gruppo di membri universitari e loro familiari di un'altra università Americana. Qui consideriamo la scala di felicità Likert a nove punti e la scala di stress Likert a sette punti

Per ricavare caratteristiche di localizzazione e mobilità significative i dati sono stati elaborati.

Il primo metodo utilizza un modello di miscela gaussiana (GMM) a tutti i campioni di localizzazione per ciascun partecipante per identificare le loro

posizioni. Si è assunto che i partecipanti frequentassero al massimo venti location, le cui principali sono la casa e il luogo di lavoro che sono stati considerati questi come il clustering completo.
Il secondo metodo utilizza il clustering K-means solo per posizioni fisse. Il dataset StudentLife includeva una previsione che l'individuo fosse o fermo in un punto o in movimento, ma l'altro dataset no.
Consideriamo questo secondo insieme come il clustering stazionario e il Cluster notturno il cluster dove ogni individuo trascorre più tempo tra le 12 PM - 6 AM.

A valle dell'identificazione dei dataset e dell'esecuzione dei metodi di preparazione dei dati sono state svolte due attività di previsione:

1. Prevedere se un individuo in un determinato giorno fosse felice/stressato o meno

2. Prevedere il livello medio di felicità o stress che un partecipante ha segnalato in un determinato giorno

Per valutare i livelli di stress o felicità in un dato giorno sono state calcolate le medie di tutte le risposte, sulla scala Likert, che un partecipante ha riportato in quel giorno. Per sapere se il partecipante è stato felice (o stressato) o meno, è stata definita una soglia sulla media giornaliera su un valore per distinguere quando gli studenti riportassero una situazione di stress o meno. Per il dataset StudentLife la soglia è "un pò stressato", per Friends and Family, si utilizza il valore medio della scala Likert come soglia.

I Risultati ottenuti da questo esempio evidenziano come siano importanti le baseline nell'apprendimento automatico in Medicina e come vengono utilizzate nella pratica.

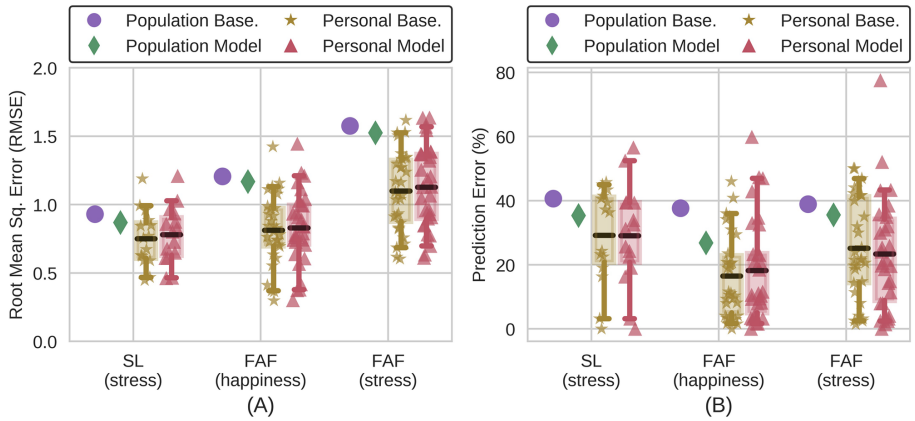

Figura 3.7: Risultati ottenuti

La figura 3.7 mostra come il modello personale produca un errore minore rispetto quello della popolazione e della baseline della popolazione, ma non minore alla baseline personale, quindi gli algoritmi predicono semplicemente gli stati riportati più frequentemente.

Per evitare il confronto con la baseline sbagliata, si propone di utilizzare test statistici con la metrica dell'user lift per dimostrare che un algoritmo sta facendo molto meglio della baseline personale.
La tecnica user lift calcola la differenza del modello personale con la baseline personale. User lift positivi indicano che un modello è migliore rispetto alla baseline personale, che le previsioni dell'algoritmo sono più accurate di quanto si supponga sempre che un individuo sia al suo stato medio.

Dataset	Problem	Model	Avg. Personal Baseline Error	Avg. Personal Model Error	Avg. User Lift (Error)	p-value
SL—Stress	binary	Log.Reg.	29.19%	29.09%	**0.10**	.481
FaF—Happiness	binary	SVM(rbf)	16.51%	18.67%	**-2.17**	.967
FaF—Stress	binary	SVM(rbf)	25.17%	23.35%	**1.82**	.240
SL—Stress	regression	Elastic Net	0.75	0.78	**-0.03**	.988
FaF—Happiness	regression	Elastic Net	0.81	0.83	**-0.02**	.999
FaF—Stress	regression	Elastic Net	1.10	1.13	**-0.03**	1.000

Figura 3.8: Risultati con tecnica user lift

La figura 3.8 mostra che gli user lift dell'utente non sono significativamente più grandi di zero e quindi i modelli non forniscono risultati migliori rispetto a delle linee di base personali costanti.

In questo studio abbiamo mostrato, con esempi di stress e felicità utilizzando due dataset pubblici, quanto sia facile che gli algoritmi di ML risultino promettenti se confrontati con le baseline non giuste. Gli individui riportano a sorpresa piccole variazioni del proprio stato mentale, quindi predire sempre che un individuo è nel suo stato d'animo più frequente è corretto nella maggioranza dei casi. Come risultato quando un algoritmo viene confrontato con una baseline della popolazione che sempre prevede che tutte le persone siano allo stesso stato, i risultati dell'algoritmo possono sembrare accurati anche se non migliorativi nel predire che ogni individuo sia nel suo stato più frequente.
È stato riscontrato che nel 77% delle pubblicazioni viene utilizzata la baseline della popolazione, e che quando viene utilizzata quella personale il risultato non cambia più di tanto, anzi a volte peggiora.
Una limitazione sul set di dati scelto è stata che i gruppi di studio non erano popolazioni cliniche, la dimensione del campione era esigua e la durata dello studio era limitata.
Mentre il target della popolazione solitamente preso in esame dovrebbero essere individui affetti con il disturbo dell'umore, i gruppi di studio presi in considerazione sono spesso di piccole dimensioni e sani mentalmente.
È possibile che le persone con disturbi dell'umore riportino più cambiamenti di stato rispetto alla popolazione normale. Una maggiore variabilità ridurrebbe la probabilità di risultati falsamente ottimistici.

La capacità di predire segnali personali significativi per il monitoraggio medico, come il benessere mentale, potrebbe migliorare notevolmente l'assistenza personalizzata consentendo approcci just in time e interventi personalizzati. Tuttavia sono state evidenziate alcune insidie nella valutazione degli algo-

ritmi per questa applicazione che possono condurre a risultati falsamente ottimistici fornendo un ottimismo infondato.
Per limitarli è stato proposto di utilizzare la metrica user lift, che si concentra direttamente sull'individuo.

Conclusioni

La trasformazione digitale della società ormai da anni sta caratterizzando la vita dei singoli individui e delle loro attività professionali. In particolare gli aspetti più innovativi di questa trasformazione sono rappresentati dalla progressiva introduzione dell'Intelligenza Artificiale nella nostra vita quotidiana e di quanto gli algoritmi di ML possano aiutarci a semplificarla.
Queste tecnologie si sono ormai affermate in tutti i settori aziendali, sia produttivi che di servizi, e tra questi spicca il settore della Medicina, dove si ripone una grande speranza nel contributo che potranno apportare nei prossimi anni all'assistenza sanitaria e più in generale alla qualità della vita dell'individuo.

L'obbiettivo della tesi è stato quello di analizzare l'introduzione dell'Intelligenza Artificiale e degli algoritmi di Machine Learning in Medicina, cercando di sottolineare non solo i risultati positivi già oggi raggiunti e le promettenti prospettive future, ma di individuare anche quelli che sono gli aspetti negativi e le perplessità legati all'introduzione di questi sistemi come supporto decisionale al personale medico nello svolgimento delle proprie attività.

Dopo una parte iniziale, in cui sono stati esposti i principi base dell'Intelligenza Artificiale, in particolare le caratteristiche che ci consentono di definire una macchina intelligente, sono state trattate le tecnologie che, grazie allo sviluppo di specifici algoritmi, rendono tali macchine "intelligenti", ovvero capaci di sviluppare un'attività "umana" senza il supporto dell'uomo.

La parte centrale di questo lavoro di tesi si è focalizzata sulle attuali criticità e perplessità legate all'Intelligenza Artificiale e Machine Learning in Medicina, ovvero sulla possibilità di non avere un reale e preciso supporto decisionale all'attività sanitaria, ma al contrario di ottenere risultati imprecisi o addirittura fuorvianti.
Analizzando la letteratura disponibile, sono stati individuati tre articoli che descrivono casi concreti in cui si dimostra come determinati algoritmi di ML possano portare a valutazioni incorrette e più in generale a criticità sul personale medico.

Nel primo caso si è evidenziato la difficoltà di dare una rappresentazione informatica completa al contesto clinico, che per sua natura è molto articolato. Questa mancanza all'interno degli algoritmi di apprendimento automatico può influenzarne negativamente il funzionamento e quindi produrre risultati non corretti e addirittura fuorvianti nel processo decisionale del medico.
Un altro elemento di criticità evidenziato è il tema dell'over-reliance, ovvero come l'introduzione del supporto decisionale dei sistemi di ML possa portare nel tempo ad un'eccessiva dipendenza da parte dei medici verso queste tecnologie, e quindi come conseguenza principale la dequalificazione (deskilling) del personale sanitario.

Nel secondo caso si è valutato come l'analisi dell'utilizzo dei social network, per esempio Facebook, da parte della comunità dei malati di morbo di Crohn possa essere di ausilio all'attività medica. La conclusione a cui si è arrivati è che ad oggi i social network non possono consentire deduzioni scientificamente valide per capire cosa stia accadendo ad un paziente, in modo particolare, a causa dell'incompletezza delle informazioni.

Nel terzo caso si è evidenziato l'importanza della verifica dell'accuratezza degli algoritmi di ML, che si può ottenere grazie al procedimento della con-

valida incrociata (CV - Cross Validation). Inoltre si è dimostrato come la scelta del metodo di CV debba essere strettamente legata allo scenario d'uso, per non indurre ad una sovrastima dell'accuratezza dell'algoritmo.

Nel quarto caso si è evidenziato come la scelta del corretto dataset sia fondamentale per non produrre risultati fuorvianti, in particolare si è dimostarto come il modello possa risultare impreciso se si confronta un singolo individuo con la baseline dell'intera popolazione.

A conclusione, occorre sottolineare come queste tecnologie possano dare un enorme contributo nell'attività medica (ad esempio processando enormi quantità di dati in tempi rapidissimi, attività impossibile per il cervello umano), ma allo stesso tempo non possono prescindere dal contributo umano del medico, che comunque deve sempre dare un senso finale ai dati analizzati anche in virtù della natura non deterministica della scienza medica.

Bibliografia

[1] https://jamanetwork.com/journals/jama/article-abstract/2645762?utm_campaign=articlePDF&utm_medium=articlePDFlink&utm_source=articlePDF&utm_content=jama.2017.7797

[2] https://link.springer.com/epdf/10.1007/s13721-016-0122-9?author_access_token=xCxIY3Z1hM3Mv3wxumRJsfe4RwlQNchNByi7wbcMAY4URozRRx0maatoXhFqHFJCjvtYGb1eP-yH9miM99MkVWHT_cnTap0KAYqAnT1ttqjlts4eqXXB5yMhhOZpbI5j_xtcCqropoTVUVKE_gJM-A%3D%3D

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5441396/

[4] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184604

[5] https://www.ai4business.it/intelligenza-artificiale/intelligenza-artificiale-cose/

[6] https://en.wikipedia.org/wiki/Machine_learning

[7] https://en.wikipedia.org/wiki/Deep_learning

[8] https://www.domedica.com/intelligenza-artificiale-e-medicina-tra-futurismo-e-concrete-realta/

[9] http://www.toscanamedica.org/89-toscana-medica/ricerca-e-clinica/613-intelligenza-artificiale-in-medicina-tra-hype-incertezza-e-scatole-nere