1

Capitolo J43
intelligenza artificiale - campi

Contenuti delle sezioni

68 pagine

J43 a. emergere della AI

J43 a.01 Sin dai primi anni della disponibilità di apparecchiature elettroniche programmabili in grado di effettuare eleborazioni automatiche si sono poste domande su quali generi di attività potevano sviluppare proficuamente e sui limiti delle loro prestazioni.

Tra i primi a porsi questo genere di domande troviamo personalità come Alan Turing, Claude Shannon, McCulloch e Pitts, Marvin Minsky e John McCarthy.

Turing nel 1936 aveva proposto il modello di macchina che porta il suo nome in particolare la cosiddetta macchina universale in grado di simulare tutte le macchine (tra le quali se stessa) giungendo quindi, con Alonzo Church a congetturare che tutte le elaborazioni di informazioni, sia numeriche che simboliche (e quindi anche le logiche) che possono essere effettuate affidabilmente con automatismi che si presumono del tutto precisi (oppure da esecutori umani perfettamente meticolosi), possono essere eseguite da una macchina di Turing.
Va segnalato anche che nel solco della ricerca sui fondamenti della matematica sono stati studiati vari altri generi di meccanismi per manipolazioni formali che si sono dimostrati equivalenti alle macchine di Turing.

La congettura di Church-Turing è ampiamente condivisa e costituisce un robusto sostegno al carattere unitario di tutte le elaborazioni che vengono effettuate seguendo regole condivisibilmente giudicate ben definite o da operatori umani meticolosi o da automatismi riconosciuti come altamente precisi, interamente controllabili e riproducibili.

Turing nel corso della II guerra mondiale nel centro britannico di decrittazione di Blatchley Park (1942-44) aveva guidato la costruzione delle cosiddette "bombs", macchine in grado di interpretare comunicazioni radiotelegrafiche degli eserciti nazisti e aveva contribuito alla progettazione e alla costruzione di alcuni dei primi computers.

Nel 1943 McCulloch e Pitts avevano proposto il modello delle reti neurali artificiali come automatismi equivalenti alle macchine di Turing ispirati dai sistemi biologici dei neuroni cerebrali.

Claude Shannon, anch'egli crittografo durante la guerra, negli ultimi anni 1940 ha posto le basi della teoria dell'informazione, o più precisamente della teoria matematica della comunicazione delle informazioni.

J43 a.02 La prospettiva di realizzare automatismi con elevate prestazioni computazionali e con la possibilità di affrontare importanti problemi ha indotto alcuni studiosi di queste macchine a porsi in forma pressante domande sulla portata delle prestazioni di questi automatismi macchine e sull'orizzonte applicativo che si poteva intravvedere.

Nel 1950 Marvin Minsky e Dean Edmonds, hanno realizzata la prima rete neurale artificiale, chiamata SNARC.

Si è inoltre iniziato a discutere il problema di cosa si possa intendere per intelligenza da considerare una capacità posseduta non solo da esseri viventi, ma anche, probabilmente, da sistemi artificiali dotati di elevata versatilità.

Nel 1950 Turing ha spostato l'attenzione dalla domanda sull'essenza dell'intelligenza (giudicata troppo vaga e lontana da strade risolutive) proponendo di esaminare il problema del riconoscimento della differenza tra le attività verbali attribuibili all'intelligenza umana e attività similmente verbali che possano essere effettuate da automatismi per in una contesa chiamata simulation game.

Questo gioco, ora noto come "test di Turing", vede un interrogatore porre domande a un umano e a una macchina nascosti per cercare di capire dalle loro risposte verbali quale sia l'umano e quale l'automatismo.
L'importanza di questo test sta nel fatto che invita a individuare casi concreti di macchine che effettuano operazioni intelligenti e contemporaneamente induce a esaminare definite operazioni eseguibili dall'intelligenza umana, riconosciuta come poco nota.

Il problema delle prestazioni intelligenti è stato affrontato anche nell'ambito della psicologia e della linguistica, soprattutto da Skinner forte dei risultati delle indagini del comportamentismo e da Chomsky rivolto allo strutturalismo per spiegare le potenzialità linguistiche dei bambini e la definizione di linguaggi artificiali che si ritenevano necessari per potenziare la usabilità degli elaboratori elettronici per effettuare calcoli automatici.

Nel 1955 Noam Chomsky ha introdotto le grammatiche generative, meccanismi per la riscrittura simbolica in grado di produrre illimitatamente le stringhe di un linguaggio formale, nella prospettiva di individuare grammatiche che potessero rappresentare sia i linguaggi naturali che i linguaggi artificiali.

Chomsky ha definito una gerarchia di quattro collezioni di grammatiche generatrici di linguaggi e quindi una gerarchia di quattro collezioni di linguaggi egli ha inoltre definite quattro collezioni di automi in grado di riconoscere le stringhe appartenenti ai linguaggi di ciascuna delle quattro collezioni di linguaggi, automi con capacità di definizione di linguaggi del tutto equivalenti alle grammatiche delle rispettive quattro collezioni.
È quindi stata precisata la cosiddetta gerarchia di Chomsky costituita da quattro collezioni di linguaggi formali via via più estese

i particolari linguaggi razionali (chiamati più comunemente "linguaggi regolari"), controllabili mediante digrafi

i linguaggi acontestuali, trattabili attraverso strutture ad albero,

linguaggi contestuali controllabili con automi sempre in grado di decidere,

linguaggi generali controllabili con macchine di Turing riconoscitrici per le quali rimane apero il problema dell'arresto.

Questa gerarchia ha consentito di distinguere con relativa semplicità quattro livelli di possibilità espressive e operative dei linguaggi formali.

J43 a.03 Nel 1956 presso il Dartmouth College, nel New Hampshire, si è tenuto un convegno dedicato allo sviluppo di automatismi che si potessero considerare intelligenti, convegno che ha visto la partecipazione di John McCarthy, Marvin Minsky e Claude Shannon. In questa sede si è esaminato ed apprezzato il programma di Allen Newell e Herbert Simon chiamato Logic Theorist in grado di dimostrare teoremi matematici.
Inoltre in questa sede si è accettata la proposta di McCarthy di adottare il termine "intelligenza artificiale".

Il convegno di Dartmouth si è concluso con molto ottimismo e con la proposta di progettare nuove macchine in grado di simulare attività che avevano sempre richiesta l'intelligenza umana.

Nel periodo che si chiude all'incirca nel 1965 nei confronti della AI si sono create grandi aspettative.
Notevoli successi sono stati ottenuti da programmi capaci di ragionamenti riguardanti problemi circoscritti, in particolare con strumenti in grado di effettuare attività logico deduttive, ossia di condurre esecuzioni inferenziali.

Newell e Simon hanno realizzato un programma chiamato General Problem Solver in grado di imitare il modo di operare di un umano per procedere euristicamente dai procedimenti e dalle conoscenze disponibili verso il raggiungimento di nuovi obiettivi.

McCarthy ha descritto un programma ideale chiamato Advice Taker in grado di dimostrare teoremi di geometria; inoltre egli ha proposto il linguaggio Lisp che per molti anni sarà il più utilizzato per affrontare problemi nell'ambito AI e in particolare i cosiddetti problemi concernenti i micro mondi, problemi risolvibili utilizzando meccanismi deduttivi.

Nel 1963 James Slagle ha definito il programma SAINT in grado di effettuare calcoli di integrali ottenibili in forma chiusa, avviando un filone di realizzazioni a supporto delle attività matematiche.

J43 a.04 Nel 1966 Weizenbaum ha realizzato il programma Eliza finalizzato a sostenere interazioni uomo-macchina attraverso il linguaggio naturale. Si è trattato di un'esperimento stimolante, ma con limiti applicativi individuati con lucidità dallo stesso proponente.
Eliza si basava sul confronto della frase presentata dell'interlocutore umano con un elenco di schemi di frase e trovato il corrispondente lo traduceva nella risposta ottenuta adattando lo schema di risposta previsto nello stesso elenco.

Da queste attività è emersa la necessità di definire e mettere a punto un esteso insieme di nuovi strumenti per la schematizzazione formale dei significati e per disporre di meccanismi efficienti di Information retrieval.

In seguito all'alternarsi di risultati promettenti e di ostacoli difficili da superare, sono stati delineati un programma "AI debole", più realistico, volto al conseguimento di risultati che rispondono ad esigenze concrete circoscritte e che prospettano soluzioni di problemi sentiti e un programma "AI forte" più ambizioso e che sostiene l'opportunità di concentrarsi su obiettivi il cui conseguimento apra la possibilità di successi della più ampia portata.

In quegli anni su questi due atteggiamenti si è sviluppato un dibattito filosofico-tecnologico spesso condotto con toni molto aspri, anche in relazione al problema delle priorità dei finanziamenti disponibili.

J43 a.05 Le prospettive dei sostenitori della AI intorno al 1965 ha incontrato molte delusioni.

Si devono tenere presenti varie affermazioni ottimistiche dei sostenitori della AI che si sono rivelate ben poco fondate.

Nel 1965 Herbert Simon aveva inoltre previsto che nel giro di 20 anni le macchine sarebbero state in grado di svolgere gran parte delle azioni umane.
Simon nel 1957 aveva stimato che nel giro di dieci anni sarebbero state disponibili macchine in grado di competere con i campioni di scacchi e di dimostrare importanti teoremi matematici.
La previsione sugli scacchi si avvererà, ma quarant'anni dopo, inseguito ad una serie di attività molto impegnative che hanno richiesto sia la crescita degli strumenti materiali disponibili, sia l'approfondimento delle conoscenze logiche, matematiche e algoritmiche, sia la maggiore disponibilità degli ambienti industriali e accademici.

Nel 1970 Minsky aveva previsto che nel giro di 3 o 8 anni sarebbe stata realizzata una macchina con l'intelligenza di una persona di media capacità mentale.

Uno dei maggiori obiettivi dei sostenitori dalla AI fin dagli anni 1950 era quello della traduzione automatica, ma fino a tutti gli anni 1960 i risultati sono stati deludenti.
Un analogo scarso successo ha avuto l'obiettivo della comprensione automatica dei testi.

In gran parte queste previsioni avevano ingenuamente sottovalutate le risorse computazionali che si devono impiegare per realizzare automatismi che devono affrontare problemi reali con tutti i loro

aspetti concreti, risorse che in quel periodo non si sapevanpo valutare correttamente e in effetti erano ancora lontane dall'essere disponibili.

J43 a.06 Il matematico James Lighthill con un rapporto scritto nel 1973 per il British Science Research Council (BSRC) ha espresso il suo pessimismo sulle possibilità della AI nei settori della robotica e dalla elaborazione dei linguaggi naturali. Egli si dichiarava molto critico sulle possibilità di una ricerca di base volta a fornire solidi principi per la AI e segnalava, molto realisticamente, gli effetti della cosiddetta esplosione combinatoria che spesso si riscontra quando si affronta la risoluzione di problemi concreti.

Il rapporto di Lighthill ha convinto il BSRC a ridurre i fondi per la ricerca sulla AI nel Regno Unito e poco dopo ha convinto il DARPA e il governo USA della ridotta importanza che la AI avrebbe avuta in quegli anni e questo ha portato alla riduzione dei fondi anche negli USA.
La conseguente situazione di crisi delle attività AI che si è venuta a creare è stata chiamata "AI winter".

J43 a.07 Le critiche all'ottimismo di buona parte dell'area AI sono state sottovalutate soprattutto da molti sostenitori del programma AI forte che di conseguenza sono stati accusati di poca professionalità.

Va detto in particolare che lo stesso Weizenbaum aveva criticato l'utilizzo del suo programma ELIZA per interventi psicoterapeutici e aveva segnalato con chiarezza che l'utilizzo ingiustificato, privo di basi pragmatiche serie, di realizzazioni AI poteva essere pesantemente pericoloso.

Mirsky e Papert nel 1969 hanno pubblicato il libro "Perceptrons", testo che esamina il progetto Perceptron di Frank Rosenblatt che aveva sostenuta l'opportunità di dare prospettive alle reti neurali; nel suddetto libro si segnalavano anche le correnti limitazioni di queste reti e questo ha portato all'abbandono delle ricerche su sistemi ispirati dal connessionismo per una decina di anni.

Un importante passo in avanti per lo sviluppo dell'area AI è consistito nella adozione di precise tecniche formali che consentono di avvalersi con profitto della logica. A questo si è rivolta la proposta di J. Alan Robinson nel 1963 di implementare queste manovre nei programmi incaricati di procedimenti deduttivi. Si è però constatato che procedendo in questa direzione si incontravano tempi di calcolo in gran parte insostenibili dalla tecnologia dei processori allora in produzione.

Maggiori successi dell'approccio che privilegia procedimenti inferenziali suggeriti dalla logica nella direzione si sono ottenuti in seguito alla implementazione del sistema per la programmazione nel linguaggio per elaborazioni logiche Prolog ispitata da Robert Kowalski e realizzata da Alain Colmerauer; questo sistema che dava il vantaggio della elevata versatilità in campi nei quali erano necessarie sperimentazioni estese e accurate.

Una debolezza dell'approccio che privilegia l'azione degli algoritmi logici deriva dal fatto che gli umani quando affrontano i problemi raramente usano la logica, ma viceversa prendono decisioni fortemente influenzate da reazioni emotive, da condizionamenti ambientali, da stereotipi e da altre propensioni che devono essere studiati con strumenti della psicologia.

J43 a.08 Nei primi anni 1970 un rilevante sostegno alle prospettive dell'intelligenza artificiale di AI è giunto dal Giappone che per sostenere la sua industria manufatturiera ha deciso di investire importanti risorse nello sviluppo della sua strumentazione digitale puntando in particolare sulla robotica.

L'università di Waseda nel 1972 ha completato la realizzazione di WABOT-1, il primo androide (ossia robot umanoide) in grado di muoversi e di spostare oggetti grazie a occhi artificiali, sensori tattili e mani prensili; esso operava interagendo con un suo controllore umano mediante conversazioni in giapponese.

Tuttavia anche per procedere sistematicamente in questa direzione è stato necessario aspettare il potenziamento degli strumenti hardware, evoluzione che si è realizzata, ineluttabilmente, ma gradualmente, in tempi necessariamente non brevi.

Per la progressione delle prestazioni dei dispositivi e dei sistemi elettronici è opportuno tenere presente la cosiddetta legge di Gordon Moore, legge formulata nel 1965 che prevedeva il raddoppio del numero dei transistors collocabili nei microprocessori e nei circuiti elettronici nel giro di 18 mesi, raddoppio che a sua volta assicurava un simile tasso crescita per la potenza dei dispositivi elettronici che riescono a entrare nel mercato.

Questa legge empirica ha indirizzato gli obiettivi della ricerca e dello sviluppo nell'industria dei componenti elettronici e si è dimostrata valida fino a circa il 2010 quando si sono raggiunti i livelli di affollamento dei transistors delle distanze di pochi nanometri.
A tali valori si riscontra la crescita degli effetti parassiti di natura quantistica e delle possibilità di interferenze da raggi cosmici, da eventi termici e da disturbi meccanici, con la conseguente richiesta di nuovi generi di soluzioni.

Di queste innovazioni diremo in seguito parlando delle cosiddette nanotecnologie. Al momento ci limitiamo a sottolineare che il ritmo di crescita della componentistica ha condizionato gli avanzamenti della gran parte delle applicazioni dell'ICT e in particolare in progressi della AI.

J43 b. Agenti

J43 b.00 Le attività AI si servono di operatori artificiali che basano le loro scelte sulle conoscenze in loro possesso; questi sono chiamati spesso "agenti logici", ma qui li chiameremo semplicemente **agenti**.

Un tale agente si esprime con un linguaggio formale che si serve di un lessico ben definito, segue precise regole sintattiche, fa riferimento a una semantica che determina la verità delle formule rispetto a un mondo possibile che si adegua al genere delle applicazioni che il linguaggio cerca di esprimere.

J43 b.01 La AI si può definire anche come la progettazione di agenti che quando hanno successo si possono giudicare intelligenti.

Genericamente diciamo agente ogni entità in grado di percepire l'ambiente nel quale opera, ovvero il workspace che lo circonda, servendosi di sensori e inoltre in grado di agire su di esso attraverso suoi attuatori.

Per definire gli scopi della AI è opportuno avere ben chiara la tipologia dei sensori e degli attuatori.

Si definisce come funzione risposta di un agente o funzione agente la corrispondenza univoca che a una sequenza di percezioni associa l'azione che l'agente intraprende come sua reazione.
Si definisce come misura di prestazione di un agente una valutazione del suo comportamento in un suo ambiente operativo.

prgCb.02 Si giudica agente razionale un agente che riesce a massimizzare la misura di prestazione corrispondente alla sequenza percettiva che ha acquisita nel corso delle attività svolte.

Per conoscere un agente serve conoscere in modo dettagliato i suoi ambienti operativi e le corrispondenti misure di prestazione.
Serve una precisa classificazione degli ambienti operativi che vanno distinti tra i completamente osservabili e i parzialmente osservabili, tra ambienti di singolo agente e ambienti multiagente, tra deterministici e non deterministici, tra episodici e sequenziali, tra statici e dinamici, tra discreti e continui e tra noti e ignoti.

prgCb.03 Nei casi di misura di prestazione ignota o incerta il profilo dell'agente deve contenere una valutazione dell'incertezza dell'obiettivo.

Per programma agente si intende l'implementazione dell'agente, ovvero della funzione agente.
Per i programmi agente vengono studiati vari schemi base che differiscono in efficienza, compattezza e flessibilità; la scelta di uno schema implementativo dipende fortemente dalle caratteristiche dell'ambiente di lavoro.

Si distinguono:
agenti reattivi semplici che alle percezioni reagiscono con risposte dirette;
agenti reattivi basati su modello che si servono di uno stato interno per tenere traccia degli aspetti dell'ambiente non segnalati dalle percezioni correnti;
agenti basati sugli obiettivi;
agenti basati sull'utilità.

Tutti gli agenti definiti nella odierna AI possono migliorare le loro prestazioni attraverso l'apprendimento automatico.

J43 c. Sulle soluzioni mediante ricerca

J43 c.01 Molte delle decisioni che devono prendere gli agenti si basano su operazioni di ricerca entro ambienti o domini dei generi più vari.

Per lo studio dell'area AI risulta quindi conveniente avere un quadro ben definito dei generi di queste operazioni.

In tutte le ricerche si devono bilanciare il tempo da impiegare nella ricerca, la memoria da occupare, e la qualità della soluzione attesa.

Le ricerche che in linea di massima sono meglio definite sono le ricerche svolte da singoli agenti, episodiche, collocate in ambienti con caratteristiche interamente note, che sono completamente osservabili e deterministici.
In questi casi è possibile aumentare l'efficienza della ricerca a partire da conoscenze sopra il dominio tradotte in una fuzione euristica che riesce a stimare la distanza tra lo stato del dominio raggiunto e lo stato obiettivo; inoltre può essere vantaggioso precalcolare soluzioni parziali utilizzando patterns e punti di riferimento.

J43 c.02 Un problema di ricerca è definito da uno stato iniziale, l'insieme delle possibili azioni che portano a transizioni, un modello delle transizioni al quale si riferiscono le azioni, un insieme di stati obiettivo e una funzione costo delle posibili azioni.
Il dominio della ricerca è rappresentato da uno spazio degli stati e la soluzione di una ricerca può vedersi come cammino dallo stato iniziale a uno stato obiettivo.

In genere gli algoritmi per queste ricerche considerano stati e azioni come entità atomiche, prive di strutture interne.
Gli algoritmi di ricerca sono valutati in termini di complessità temporale, di complessità spaziale, di completezza e di ottimalità.

I metodi di ricerca non informata si servono degli algoritmi che seguono.
Gli algoritmi costruiscono un albero di ricerca che aiuti a trovare un cammino rappresentante lo svolgimento di una soluzione e si distinguono principalmente sulla base del modo di scegliere, per ogni transizione, il nodo dell'albero da esaminare per primo.

La ricerca best fit sceglie i nodi da esaminare attraverso una funzione di valutazione.

La ricerca in ampiezza esamina tutti i nodi figlio del nodo attualmente raggiunto; quando le azioni hanno lo stesso costo questa ricerca è ottima, ma ha la complessità spaziale con andamento esponenziale.

La ricerca a costo uniforme sceglie il nodo figlio dal costo minore ed è ottima per ogni costo di passo.

La ricerca in profondità espande per primo il nodo più profondo non ancora espanso, non è ottima e non è completa, ma presenta una complessità spaziale lineare, decisamente conveniente.

Viene adottata anche una variante della precedente detta ricerca in profondità limitata.

La ricerca ad approfondimento iterativo esegue ricerche a profondità limitata estendendo progressivamente il limite della profondità.
Si ottiene una ricerca completa quando si effettua un controllo completo su tutti i cicli; questa presenta complessità temporale comparabile con la ricerca in ampiezza e presenta complessità spaziale lineare.

Viene adottata anche la ricerca bidirezionale che opers su due frontiere di stati esaminati, una ottenuta a partire dello stato iniziale e una a partire dallo stato obiettivo.

J43 c.03 I metodi di ricerca informata hanno a disposizione una funzione euristica $h(n)$ che stima un costo della soluzione a partire dal nodo n.
Talora la ricerca informata dispone anche di una base di dati di patterns con i costi delle soluzioni raggiungibili. Vediamo i tipi principali di questo genere di algoritmi.

La ricerca best-first greedy: espande per primi i nodi con $h(n)$ minima; non è ottima, ma spesso si rivela efficiente.

La ricerca A^* espande i nodi con $f(n) := g(n) + h(n)$ minima, è completa e ottima quando $h(n)$ è ammissibile; purtroppo per molti problemi la complessità spaziale cresce notevolmente e questo costituisce una difficoltà.

La ricerca IDA^* (iterative deepening A^*) è una variante della precedente ad approfondimento relativo che affronta il problema della complessità spaziale in misura attenuata.

J43 c.04 La ricerca RBFS (Recursive best fit search) e la ricerca SMA^* (simplified memory bounded A^*) si rivelano robuste, ottime e con utilizzo di una quantità di memoria limitata.
Disponendo di tempi sufficienti questo metodo risolve problemi che il metodo A^* non giunge a risolvere in quanto esaurisce la memoria.

La ricerca beam pone un limite alla estensione della frontiera degli stati esaminati e questo la rende incompleta e subottima; tuttavia essa spesso trova soluzioni ragionevolmente buone con tempi sensibilmente inferiori di quelli richiesti dalle ricerche più esaurienti.

La ricerca A^* pesata fa tendere il cammino risolutivo verso un obiettivo, espandendo meno nodi e con questo rinunciando all'ottimalità.

Le prestazioni i questi algoritmi dipendono sensibilmente dalla qualità della funzione euristica e talvolta si riescono ad ottenere euristiche vantaggiose rilassando gli obiettivi del problema, utilizzando collezioni di patterns con costi precalcolati delle soluzioni, definendo punti di riferimento e facendosi orientare dall'esperienza.

J43 c.05 Si utilizzano algoritmi di ricerca anche per ambienti parzialmente osservabili, non deterministici e rappresentati da spazi continui.

Si adottano molti metodi di ricerca locale chiamati di "hill climbing" i quali mantengono in memoria solo un piccolo insieme di stati. Essi sono stati applicati a problemi di ottimizzazione e cercano uno stato ad alto punteggio senza preoccuparsi del cammino da compiere per raggiungerlo.
Sono adottati svariati algoritmi stocastici, tra i quali quelli di simulated annealing i quali riescono a fornire soluzioni ottime quando si individuano velocità di raffreddamento adeguate.

Sono stati studiati molti metodi di ricerca locale anche per ambienti rappresentati da spazi continui.
Sono stati adottati metodi di programmazione lineare e di ottimizzazione convessa che rispettano precisi vincoli per lo spazio ambiente e per le caratteristiche della funzione obiettivo. Questi riescono a servirsi di algoritmi con complessità temporale polinomiale spesso molto efficienti per situazioni che rivestono rilevante importanza pratica.

Vi sono anche problemi che possono essere impostati con formalismi matematici maneggevoli per i quali dove il gradiente è zero è possibile procedere con operazioni analitiche; in altri casi si deve invece ricorrere a gradienti ottenuti empiricamente dalle differenze di fitness tra punti vicini.

Sono usati algoritmi evolutivi di ricerca hill climbing stocastici che tengono in memoria intere popolazioni di stati e procedono a individuare nuovi stati con operazioni di mutazione e di crossover su coppie di stati.

J43 c.06 In ambienti non deterministici gli agenti possono effettuare ricerche AND-OR per determinare piani condizionali in grado di far raggiungere l'obiettivo quali che siano i risultati ottenuti nel corso della esecuzione.

Nel caso di ambiente parzialmente osservabile, si fa riferimento al cosiddetto "stato credenza", entità introdotta con il ruolo del rappresentante dell'insieme degli stati nei quali l'agente potrebbe trovarsi vantaggiosamente.

Per risolvere problemi nei quali l'agente non può usare sensori vengono adottati i vari algoritmi standard direttamente allo spazio degli stati credenza; una ricerca AND-OR su questo spazio può risolvere problemi parzialmente osservabili di portata generale.
Spesso invece risultano più efficienti algoritmi incrementali che costruiscono soluzioni stato per stato muovendosi all'interno di uno stato credenza.

Quando un agente non conosce gli stati e le azioni dell'ambiente nel quale viene collocato si pongono problemi di esplorazione.
Se l'ambiente è esplorabile con sicurezza gli agenti possono svolgere ricerche online per costruire una mappa sulla quale cercare un obiettivo.
Per sfuggire ai minimi locali può risultare efficace tenere aggiornate stime euristiche con risultati ottenuti in precedenza.

J43 d. Ricerca in presenza di avversari

J43 d.01 Molti problemi affrontati dalla AI riguardano ambienti competitivi nei quali operano due o più agenti con obiettivi in conflitto; per questi problemi si rendono necessarie le cosiddette operazioni di ricerca con avversari.

Gli studiosi di AI si sono occupati soprattutto di giochi come dama, scacchi, Go e poker; per questi si ha un campo di gioco facile da rappresentare in quanto gli agenti possono effettuare solo azioni ben definite con effetti che si può cercare di valutare con precisione.
La realizzazione di programmi in grado di battere avversari umani qualificati in competizioni appositamente organizzate e publicizzate ha contribuito a consolidare la fiducia nelle possibilità delle elaborazioni studiate dalla AI. Le esperienze maturate hanno portato a una visione via via più ampia e consolidata dei meccanismi che gli agenti AI possono mettere in campo nelle varie applicazioni nelle quali si trovano a competere con avversari.

J43 d.02 Un gioco è definito dallo stato iniziale corrispondente a una configurazione iniziale del tavolo da gioco, dalle azioni che possono essere messe in atto nei vari stati, dalle conseguenze dei diversi tipi di azioni, da un test di conclusione del gioco e da una funzione di utilità delle configurazioni che alla fine di una partita decidono i punteggi conclusivi e un conseguente vincitore.

Consideriamo i giochi a due giocatori, discreti, deterministici, a turni, a somma zero e con informazione perfetta (come gli scacchi).
In tali giochi l'algoritmo minimax sceglie per un agente giocatore le mosse preferibili (possibilmente le ottime) esaminando l'albero del gioco fino alla maggiore profondità che riesce a raggiungere.

Alternativamente l'algoritmo di ricerca alfa-beta può ottenere le stesse mosse ottime di minimax, ma ha maggiore efficienza perché sa trascurare sottoalberi do gioco giudicati irrilevanti.

In genere non si riesce a esaminare l'intero albero di gioco, anche a meno dei sottoalberi che alfa-beta sa scartare; quindi bisogna ridurre le operazioni di ricerca in genere servendosi di una qualche funzione euristica che fornisca una valutazione della rilevanza dei nodi dell'albero.

J43 d.03 La cosiddetta ricerca dell'albero Monte Carlo valuta gli stati simulando con una certa aleatorietà proseguimenti di partita fino alla conclusione in modo da capire chi potrà vincere. Data la aletorietà delle mosse scelte in una simulazione conviene effettuare più simulazioni e assumere una valutazione media per ogni nodo in esame.

Molti programmi di gioco si servono di sequenze precalcolate di mosse migliori per le fasi iniziali e finali delle partite in modo da poterle usare come scelte, rinunciando a effettuare una ricerca del tutto autonoma.

I giochi che prevedono eventi casuali possono essere affrontati con la funzione expectiminimax, estensione della minimax che valuta un nodo soggetto a casualità assumendo l'utilità media sui nodi suoi successori pesata con una probabilità attribuita a ciascuno di essi.

Nei giochi con informazione incompleta, come il poker e molti altri giochi di carte, per giocare in modo ottimo bisogna esaminare i cosiddetti stati credenza individuati in un futuro prossimo prevedibile per i vari giocatori. Valutazioni approssimate si possono ottenere con la media dei valori delle azioni per tutte le configurazioni dell'informazione mancante.

I programi AI che adottano i metodi accennati con gli avanzamenti delle tecnologie computazionali hanno nettamente superati i campioni umani nella massima parte dei giochi per i quali si sono organizzate sfide palesi: dama, scacchi, Go, poker, Othello, ... I migliori giocatori umani mantengono la superiorità in pochi giochi con informazione imperfette come bridge e Kriegspiel.

In molti videogiochi i programmi sono competitivi rispetto agli umani più forti; verosimilmente questo è dovuto alla loro superiore rapidità nell'esecuzione di molte e lunghe sequenze di azioni.

J43 e. Problemi di soddisfacimento di vincoli

J43 e.01 Questi problemi, denotati anche dalla sigla CSP, constraint satisfaction problems, riguardano ricerche in ambienti i cui stati non hanno consistenza atomica, ma presentano elementi di differenziazione espressi dalle cosiddette rappresentazioni fattorizzate costituite da valori assegnati a un insieme di (operativamente una sequenza) di variabili.
Le soluzioni di questi problemi devono soddisfare una sequenza di vincoli per i valori delle variabili.

A questa categoria afferiscono molti problemi di interesse pratico; esempi: colorazione di mappe, job-shop scheduling ossia programmazione dei lavori nelle fabbriche e in altri ambienti produttivi evidentemente soggetti a vincoli di disponibilità, definizione di orari per corsi di studi, sudoku, sue varianti e analoghi rompicapo.
Spesso possono essere adottate tecniche di inferenza che dai vincoli ricavano l'esclusione di alcuni valori assegnabili alle variabili; tra queste le tecniche di consistenza di un nodo, di un arco o di un cammino e le tecniche di k-consistenza.

J43 e.02 Per risolvere i CSP, i problemi di rispetto di vincoli, comunemente si usa una ricerca con backtracking, ossia una ricerca in una certa profondità; inoltre si possono organizzare alternanze tra inferenza e ricerca.

Durante una ricerca con backtracking per scegliere quale variabile considerare tra le rimanenti si possono usare l'euristica MRV, minimum remaining values, e l'euristica di grado, entrambe indipendenti dal dominio.
Similmente l'euristica del valore meno vincolante aiuta a decidere quale valore provare per primo per una data variabile.

Il backtracking si effettua quando non è possibile assegnare un valore legale a una variabile.

Si può anche adottare il backjumping guidato dai conflitti che riporta ai primi tentativi della ricerca, ossia effettua un salto sull'albero di un buon numero di livelli.

Spesso viene adottata la manovra di apprendimento dei vincoli, la quale consiste nel registrare i conflitti che si sono incontrati per poterli evitare qualora si ripresentino.

Ricordimo anche che a molti problemi CSP si è applicata con notevole successo l'euristica min-conflicts.

J43 e.03 La complessità del processo risolutivo di un constraint satisfaction problem (CSP) dipende fortemente dalle caratteristiche del suo grafo dei vincoli. I problemi schematizzabili con un albero di scelte possono essere risolti in tempi lineari.

Un condizionamento esprimibile con insiemi di taglio può ridurre un CSP generico in uno schematizzato da un albero di scelte e se per questo si riesce a trovare un insieme di taglio piccolo si può risolvere il relativo problema con un impiego di memoria solo lineare.

Le tecniche di scomposizion ad albero trasformano il CSP in un albero di sottoproblemi e risultano efficienti se la larghezza dell'albero del grafo dei vincoli è ridotta; tuttavia queste tecniche necessitano di quantità di memoria che crescono in modo esponenziale nella larghezza dell'albero del grafo dei vincoli.

Combinando condizionamento con insiemi di taglio e scomposizioni ad albero si possono ottenere buoni bilanciamenti tra le richieste di tempo e di memoria.

J43 f. Agenti logici e inferenza nella logica del primo ordine

J43 f.01 Gli agenti che possono svolgere meglio i compiti loro assegnati sono quelli che possono meglio utilizzare le conoscenze che hanno sopra il loro spazio di lavoro.
Queste conoscenze possono essere presentate come formule espresse in un linguaggio di rappresentazione della conoscenza e possono essere registrate in una base di conoscenza.

Sa meglio utilizzare le conoscenze un agente capace di elaborare le dette formule con propri meccanismi inferenziali che consentano di ricavare nuove formule adatte a fargli decidere quali azioni intraprendere nelle varie specifiche circostanze che deve affrontare.
Un tale agente viene detto agente dotato di logica.

Un linguaggio di rappresentazione della conoscenza è definito da una sintassi che specifica la struttura delle sue formule, e da una semantica che definisce il valore di verità di ogni formula in ciascuno dei modelli degli spazi di lavoro dell'agente.

Tra le formule sussiste una relazione di conseguenza logica e con questa relazione si costruiscono i ragionamenti inferenziali dell'agente.
Da una formula α consegue una formula β se β è vera in tutti i modelli nei quali α è vera e in tal caso si scrive $\alpha \Rightarrow \beta$. Questa equivale alla insoddisfacibilità della formula $\alpha \ \wedge \ \neg\beta$.

J43 f.02 In ogni teoria formale l'inferenza è il processo mediante il quale da un insieme di formule si derivano formule che sono loro conseguenze logiche.

Algoritmi di inferenza corretti sono quelli che derivano solo formule che sono conseguenze logiche.
Si dicono invece algoritmi completi sono quelli che sono in grado di derivare tutte le formule che sono conseguenze logiche.

La logica proposizionale è un linguaggio semplice che si serve di simboli proposizionali e di connettivi logici; esso può trattare proposizioni che sono notoriamente vere, formule che sono notoriamente false e formule indefinite, che si suppone possibile dimostrare che siano vere oppure false indipendentemente da circostanze influenzabili da altre proposizioni oggetto dello studio attuale.

Consideriamo una raccolta finita o numerabile di proposizioni. Chiamo modello della raccolta ogni funzione che a ciascuna delle sue proposizioni assegna il valore vero o il valore falso.
Se la raccolta è finita l'insieme dei possibili modelli che le riguardano è finito e la loro conseguenza logica può essere verificata scorrendo questi modelli.
Tra gli algoritmi di model checking efficienti per la logica proposizionale vi sono metodi basati sul backtracking e ricerca locale che spesso possono risolvere in modo efficiente problemi di grandi dimensioni, ossia problemi con un elevato numero di proposizioni.

J43 f.03 Le regole di inferenza sono schemi di passi di ragionamento e vengono usate per costruire, scoprire e verificare dimostrazioni.
La regola di risoluzione è alla base di un algoritmo di inferenza completo applicabile a basi di conoscenza costituite da formule in forma normale congiuntiva.
Concatenazione in avanti e cocatenazione all'indietro sono algoritmi di ragionamento molto intuitivi applicabili a basi di conoscenza espresse nella cosiddetta forma di Horn.
I metodi di ricerca locale, come WALKSAT, possono essere utilizzati per trovare soluzioni, ma non sono metodi completi.

La stima dello stato attraverso la logica richiede di gestire una formula logica che esprime l'insieme dei possibili stati consistenti i risultati delle osservazioni effettuate sullo scenario chematizzato dal modello.
Ogni passo di aggiornamento richiede di effettuare una inferenza utilizzando il modello delle transizioni nell'ambiente costituito da assiomi di stato successore che determinano come cambia ogni fluente, ossia ogni condizione che può cambiare nel tempo.
Un agente logico si serve della risoluzione di un problema SAT, cioè della soluzione di un problema di soddisfacibilità, per prendere varie decisioni, ossia per trovare modelli possibili che conducono a sequenze di azioni che portano a raggiungere l'obiettivo.
Questo approccio funziona solo per ambienti completamente osservabili o quando non si devono consultare sensori.

J43 f.03 La logica proposizionale non è adatta a ambienti di elevata estensione per mancanza della portata espressiva necessaria per affrontare in modo conciso tempo, spazio e schemi generali di relazioni tra gli oggetti in esame.

La logica del primo ordine, FOL o calcolo dei predicati, è un sistema formale con una incisivitè e una portata molto superiori a quella della logica proposizionale, che qui denotiamo con PropL. Per PropL rinviamo anche a B60 e per FOL a B61.

Sul piano dell'impegno ontologico la logica proposizionale si occupa solo di fatti che possono essere solo veri o falsi, mentre la logica del primo ordine può applicarsi a campi di esistenza di oggetti e relazioni di vari generi e dispone di un potere espressivo che le consente di esaminare e influenzare una vasta gamma di situazioni problematiche.
Tra questi ricordiamo i circuiti elettronici che possono avariarsi e programmi che possono risultare scorretti.

A McCarthy si deve il chiarimnto della necessità di adottare la logica del primo ordine come strumento inferenziale dei sistemi AI.
Successivamente Cordell Green ha sviluppato un sistema di ragionamento del primo ordine adottato con successo per la robotica.
A loro volta Zohar Manna e Richard Waldingger hanno applicata la logica del primo ordine ai ragionamenti sui programmi e Michael Genesereth l'ha applicata ai circuiti elettronici.

J43 f.04 Accade però che le due logiche citate e chiamate logiche classiche non sono in grado di trattare efficacemente situazioni vaghe, con oggetti e relazioni poco definite dipendenti da giudizi personali o da altri fattori poco controllati.
Questo accade per la politica, le espressioni artistiche, la cucina e le altre questioni nelle quali possono essere determinanti gusti e pregiudizi personali e si hanno possibilità ridotte di controllo.
Inoltre le logiche classiche sono atemporali e non contingentali, non tengono conto delle variazioni dovute allo scorrere del tempo e al mutare di circostanze in grado di influenzare scelte ed eventi.

In questi casi si devono adottare logiche più elaborate.
Nelle logiche modali è possibile trattare il modo in cui viene giudicata la validità e in particolare la verità vs. falsità delle proposizioni; queste logiche sono in grado di occuparsi di questioni come possibilità, necessità, obbligo morale e credenze.

Si studiano inoltre logiche temporali che possono trattare la assegnazione alle proposizioni del valore vero o falso o in genere della validità.

J43 f.05 La sintassi della FOL estende quella della PropL aggiungendo variabili e termini che rappresentano oggetti semplici o compositi e i quantificatori esistenziale ($\exists$) e universale ($\forall$) che consentono di esprimere asserzioni sui possibili valori di verità degli oggetti e delle variabili.

Un modello o mondo possibile per la FOL è costituito da un insieme di oggetti e una interpretazione che associa simboli di costante ad oggetti, simboli di predicato a relazioni tra oggetti e simboli di funzione a funzioni sopra oggetti.

Un predicato espresso da una formula atomica è giudicato vero sse la relazione che rappresenta è verificata tra gli oggetti rappresentati dai termini.
Si trattano anche interpretazioni estese che associano le variabili dei quantificatori a oggetti del modello e stabiliscono il valore di verità delle formule quantificate.

Precisare una base di conoscenza nella logica FOL richiede un accurata analisi del dominio, una scelta di un vocabolario e la codifica degli assiomi dai quali possono essere derivate le catene inferenziali utili per le applicazioni prospettate.

J43 f.06 Per rendere operative le inferenze nella logica del primo ordine sono stati definiti vari algoritmi.

Un primo approccio consiste nell'usare regole di inferenza, ossia istanziazione universale e istanziazione esistenziale per tradurre il problema in predicati: questo modo di fare tuttavia se il dominio non è molto ridotto si rivela molto lento.
Questo approccio in molti casi viene reso più efficiente adottando l'unificazione per identificare le sostituzioni appropriate delle variabili eliminando il passo di istanziazione nelle dimostrazioni della FOL.

La regola del modus ponens può essere "sollevata attraverso il lifting" e porta a una regola di inferenza intuitiva ed efficiente chiamata **modus ponens generalizzato**.
Gli algoritmi di concatenazione in avanti e di concatenazione all'indietro si servono di questa regola e di clausole definite.

Il modus ponens generalizzato è completo per le clausole definite, ma il problema della conseguenza logica rimane semidefinito. Tuttavia la conseguenza logica è decidibile nei casi di basi di conoscenza del genere chiamato Datalog che consistono di clausole definite prive di simboli di funzione.

J43 f.07 La concatenazione in avanti è usata nei database deduttivi, in quanto può essere combinata con le classiche operazioni booleane applicate alle basi dati relazionali.

Essa è utilizzata anche nei sistemi di produzioni che eseguono aggiornamenti efficienti, anche in presenza di sistemi di regole molto estesi. La concatenazione in avanti è completa per basi di conoscenza Datalog e richiede tempi di esecuzione polinomiali.

La concatenazione all'indietro viene usata nei sistemi di programmazione logica che si servono di sofisticati compilatori in grado di ottenere inferenze molto veloci.
Tuttavia essa rischia di incorrere in problemi di interferenze e di cicli ridondanti.

Il linguaggio Prolog non riesce a controllare l'intera FOL e si rivolge a un mondo chiuso dall'ipotesi dei nomi unici e dalla negazione da valutare come fallimento. Si tratta quindi di un linguaggio di programmazione pratico, ma meno ambizioso dell'intera FOL.

La regola di inferenza della riduzione costituisce un sistema completo di dimostrazione per la FOL ed è applicabile a basi di conoscenza in forma normale congiuntiva.

J43 f.08 Sono state proposte molte strategie volte alla riduzione dello spazio di ricerca di un sistema di risoluzione senza comprometterne la completezza. Uno dei problemi più critici è la gestione

dell'uguaglianza; per questo si possono usare le tecniche chiamate "demodulazione" e "paramodulazione".

Segnaliamo anche che sono stati realizzati dimostratori di teoremi efficienti che sono riusciti a dimostrare interessanti teoremi matematici e a verificare e sintetizzare software e circuiti hardware.

J43 g. Rappresentazione della conoscenza

J43 g.01 Mentre nei casi di studio giocattolo le modalità di rappresentazione della conoscenza non hanno molto peso, per le applicazioni impegnative, quelle che presentano interesse pratico, si devono adottare rappresentazioni con elevate doti di efficienza d'uso, di flessibilità e di generalità.

Per affrontare questi problemi, ossia per rappresentare conoscenze di ampio raggio, si è costituita una disciplina detta **ingegneria ontologica**.
Essa riguarda anche una ontologia generale che organizzi e colleghi le ontologie dei diversi domini di conoscenza.

In prospettiva questi studi dovrebbero gestire una grande varietà di conoscenze; la strumentazione fisica e metodologica e le infrastrutture attualmente disponibili fanno ben sperare.

J43 g.02 Si è consolidata una cosiddetta ontologia superiore basata su categorie e calcoli degli eventi; essa sviluppa la sua visione e le sue soluzioni ed estende il suo controllo a categorie, sottocategorie, parti, oggetti strutturati, misure, sostanze, eventi, spazio, tempo, cambiamento e credenze.

I tipi naturali possono essere definiti in modo completo nella logica, ma le proprietà dei tipi naturali possono essere rappresentate .

Le azioni, gli eventi e il tempo possono essere rappresentati con il calcolo degli eventi. Queste rappresentazioni permettono a un agente di delineare sequenze di azioni e di effettuare inferenze logiche su ciò che si verificherà quando le azioni scelte saranno eseguite.

Per organizzare le categorie sono stati sviluppati sistemi di rappresentazione specializzati come le reti semantiche e le logiche descrittive.
L'ereditarietà è una forma importante di inferenza che permette di dedurre proprietà degli oggetti dalle categorie alle quali appartengono.
L'ipotesi del mondo chiuso come viene implementata nei programmi logici costituisce un modo semplice per risparmiare numerose informazioni negative. Essa si può usare come caratterizzazione di default alla quale si possono aggiungere restrizioni aggiuntive di portata specifica.

Le logiche non monotone come la circoscrizione e la logica di default costituiscono strumenti per gestire il ragionamento per default in generale. JP Sono disponibili anche sistemi per il mantenimento della verità in grado di gestire efficientemente aggiornamenti e revisioni delle conoscenze.

J43 g.03 Dato che risulta impegnativo e difficile costruire manualmente ontologie di grandi dimensioni, è opportuno aiutarsi con procedure di estrazione della conoscenza dai testi.

A questo proposito conviene ricordare che il progetto DBPedia avente lo scopo di estrarre dati strutturati da Wikipedia.
Nel 2015 si è valutato che la versione in inglese contenesse 400 milioni di fatti concernenti 4 milioni di oggetti; da tutte le versioni scritte nelle varie lingue si potrebbero estrarre 1.5 miliardi di fatti.

J43 g.04 Sono stati sviluppate altre aree per la rappresentazione della conoscenza da utilizzare per operare sopra dichiarazioni di persone che possono essere utilizzate dagli agenti AI.

La fisica qualitativa è la branca della rappresentazione della conoscenza che si propone di definire una descrizione logica e degli oggetti e dei fenomeni fisici che non sia numerica ma che abbia una sua logica.
In questo ambito sono stati resi diponibili procedimenti per trattare "storie" ambientate in sezioni del cronotopo che permettono di descrivere operazioni sui liquidi e per trattare sistemi fisici rappresentati

da astrazioni qualitative delle equazioni che li governano. Queste descrizioni sono state utilizzate per progettare dispositivi e robots per applicazioni pratiche.

Sono stati studiati modi per sviluppare ragionamenti di senso comune che consentano di trattare situazioni spaziali e di sviluppare ragionamenti spaziali qualitativi: questi hanno contribuito a defnire nuovi sistemi informativi territoriali che consentono a un agente di muoversi su regioni senza ricorrere a descrizioni metriche complete spesso non disponibili.

Si sono studiati modelli per ragionamenti su eventi psicologici ad uso degli agenti artificiali che ricorrono alla cosiddetta "psicologia popolare" utilizzata comunemente d gran parte delle persone per ragionare sui comportamenti propri e altrui.

Attualmente il maggiore utilizzo del ragionamento psicologico riguarda la comprensione del linguaggio naturale per la quale la valutazione delle intenzioni del parlante ha importanza primaria.

J43 h. Pianificazione automatica

J43 h.01 Pianificare un complesso di azioni è una competenza fondamentale per un agente artificiale e per una buona pianificazione è cruciale scegliere una buona rappresentazione dell'ambiente.
Gli studiosi della pianificazione per i procedimenti AI hanno definito un tipo di rappresentazione fattorizzata utilizzando un linguaggio chiamato PDDL, planning domain definition language (basato su Lisp) che consente di definire le azioni effettuabili da un agente attraverso un unico schema e non richiede una conoscenza specifica del dominio (V. RNi354).

Gli algoritmi di pianificazione risolvono problemi espressi mediante rappresentazioni fattorizzate esplicite degli stati e delle azioni le quali rendono possibile derivare euristiche efficaci e quindi sviluppare algoritmi potenti e flessibili.

Con PDDL si descrivono gli stati iniziali e l'obiettivo come congiunzioni di letterali e le azioni in termini di precondizioni e di effetti conseguiti.
Sono state messe a punto estensioni di PDDL che rappresentano tempo, risorse, percezioni, piani condizionali e piani gerarchici.

J43 h.02 La ricerca nello spazio degli stati può procedere in avanti (progressione) o all'indietro (regressione).
Si possono derivare euristiche efficaci ammettendo l'ipotesi della indipendenza dei sottoobiettivi oppure mediante rilassamenti delle esigenze richieste nei problemui di pianificazione.

Altri approcci si servono della codifica del problema di pianificazione come problema di soddisfacibilità booleana o come problema di soddisfacimento di vincoli e la conseguente ricerca in uno spazio opportuno di piani parzialmente ordinati.

La pianificazione basata su reti gerarchiche di compiti HTN, hyerarchical task network, consente all'agente di ricorrere a consigli del progettista del dominio espressi come azioni di alto livello (HLA, ossia high level actions) che possono essere implementate in vari modi come sequenze di azioni di livello inferiore.

Gli effetti delle HLA possono essere definiti con una cosiddetta semantica angelica che consente di derivare piani di alto livello dei quali è possibile dimostrare la correttezza senza tener conto delle implementazioni dei livelli inferiori.
Gli attuali metodi HTN facilitano la definizione dei piani molto elaborati richiesti dalle applicazioni di interesse pratico.

I piani condizionali permettono all'agente di percepire il mondo durante le sue esecuzioni e di decidere quali dei possibili rami del piano percorrere.

In taluni casi la pianificazione senza sensori, detta anche pianificazione conformante, può consentire di costruire un piano che funziona anche senza ricevere percezioni durante l'esecuzione.

Sia i piani conformanti che i piani condizionali possono essere costruiti con una ricera nello spazio degli stati-credenza.
La rappresentazione e il calcolo efficiente degli stati-credenza sono elementi fondamentali.

J43 h.03 Un agente di pianificazione online sfrutta il monitoraggio dell'esecuzione e introduce riparazioni nel piano quando si rende necessario rimediare a situazioni inattese che possono essere dovute ad azioni non deterministiche, a eventi esogeni o a modelli dell'ambiente inadeguati.

Molte azioni consumano risorse come carburante, denaro o materiali di consumo; queste risorse conviene trattarle con valutazioni numeriche forfettarie, invece che cercare di individuare e valutare i dettagli.

Il tempo, una delle risorse più importanti, può essere gestito da specifici sottoalgoritmi di scheduling; in alternativa si può integrare lo scheduling nella pianificazione.

J43 i. Quantificare l'incertezza

J43 i.01 Gli agenti che affrontano problemi reali si trovano spesso ad affrontare l'incertezza che può essere dovuta a osservabilità parziale, a non determinismo o a presenza di avversari.

Gli agenti logici gestiscono l'incertezza tenendo traccia di uno stato credenza costituente una rappresentazione dell'insieme degli stati nei quali l'agente potrebbe trovarsi e producendo un piano condizionale che gestisca ogni possibile eventualità che i suoi sensori possono percepire nel corso dell'esecuzione.

Questo approccio nei problemi riguardanti situazioni articolate può condurre a stati credenza enormi e pieni di possibilità poco probabili; può anche comportare un piano condizionale pletorico poco gestibile. Serve quindi un modo di confrontare i piani che non offrono garanzie.

J43 i.02 L'incapacità di un agente di arrivare a una decisione definitiva sul valore di verità di una formula viene espressa da una probabilità e le probabilità riguardanti le diverse formule servono a riassumere le credenze dell'agente rispetto all'evidenza.

La teoria delle decisioni combina le credenze e i desideri dell'agente e definisce come azione migliore quella che massimizza l'utilità attesa.
Gli enunciati base della probabilità includono la probabilità a priori o non condizionata e la probabilità a posteriori o condizionata riguardanti proposizioni sia semplici che complesse.
Gli assiomi del calcolo delle probabilità vincolano le probabilità di proposizioni che riguardano eventi collegati. Un agente che violasse gli assiomi in alcuni casi si comporterebbe irrazionalmente.

La distribuzione di probabilità congiunta completa specifica la probabilità delle possibili assegnazioni complete di valori alle variabili casuali.
In genere essa è troppo impegnativa da definire e da utilizzare in forma esplicita, ma quando è disponibile consente di rispondere alle varie interrogazioni semplicemente sommando gli elementi per i mondi possibili corrispondenti alle proposizioni della query attuale.

J43 i.03 L'indipendenza assoluta tra due sottoinsiemi di variabili casuali consente di fattorizzare la distribuzione congiunta completa in distribuzioni congiunte più ridotte riducendone notevolmente la complessità delle decisioni.
La regola di Bayes permette di calcolare probabilità sconosciute a partire da probabilità condizionate note, solitamente procedendo nella direzione causale.
In presenza di molti elementi di evidenza, l'applicazione della regola di Bayes ha gli stessi problemi di scalabilità dell'uso diretto della distribuzione congiunta completa.
L'indipendenza condizionale dovuta a relazioni causali dirette nel dominio permette di fattorizzare la distribuzione congiunta completa mediante distribuzioni condizionate più ridotte.
Il modello di Bayes ingenuo assume come ipotesi l'indipendenza condizionale di tutte le variabili effetto in conseguenza di una singola variabile causa e la sua dimensione cresce linearmente con il numero degli effetti.

J43 j. Ragionamento probabilistico

J43 j.01 Una rete bayesiana è un digrafo aciclico i cui nodi corrispondono a variabili casuali; a ogni nodo è associata anche una distribuzione condizionata dei suoi ascendenti.
Le reti bayesiane forniscono una rappresentazione concisa delle relazioni di indipendenza condizionale nel dominio nel quale opera l'agente.
Una rete bayesiana specifica una distribuzione di probabilità congiunta sulle sue variabili.
La probabilità di ogni assegnamento di tutte le variabili è definita come il prodotto degli elementi corrispondenti nelle distribuzioni condizionate locali.

Spesso una rete bayesiana risulta esponenzialmente più ridotta della corrispondente distribuzione congiunta presentata dettagliatamente.

Molte distribuzioni condizionate possono essere rappresentate in forma compatta facendo ricorso a famiglie canoniche di distribuzioni.
Le reti bayesiane ibride, reti che includono sia variabili discrete che continue, possono utilizzare una notevole varietà di distribuzioni canoniche.

J43 j.02 Effettuare una inferenza nelle reti bayesiane significa calcolare la distribuzione di probabilità di un insieme di variabili di query, dato un insieme di variabili di evidenza. Algoritmi di inferenza esatti, come l'eliminazione di variabili, valutano somme di prodotti di variabili condizionate nel modo più efficiente possibile.

Nelle reti unilateralmente connesse, dette anche polialberi, cioè nelle reti nelle quali per ogni coppia di nodi c'è un solo percorso non orientato dal primo al secondo,l'interferenza esatta richiede un tempo lineare nella dimensione della rete; nel caso di rete generale il problema risulta invece intrattabile.

Tecniche di campionamento casuale come la pesatura di verosimiglianza e gli algoritmi Monte Carlo per catene di Markov possono fornire stime ragionevoli delle vere probabilità a posteriori in una rete e riescono a trattare reti molto più grandi di quanto riescano a fare gli algoritmi esatti.

Mentre le reti bayesiane esprimono inferenze probabilistiche, le reti causali esprimono relazioni causali e consentono di predire gli effetti degli interventi e delle osservazioni.

J43 j.03 La mutevoleza del mondo nel quale va ambientata una problematica realistica deve essere gestita mediante un insieme di variabili casuali che contribuiscono a rappresentarlo in ogni istante.

Si ipotizza che le rappresentazioni possono essere progettate per soddisfare (almeno approssimatamente) la proprietà di Markov in modo che, dato il presente, il futuro possa essere indipendente dal passato.
Questa ipotesi insieme a quella che il processo sia omogeneo rispetto al tempo, permette di semplificare significativamente la rappresentazione della mutevolezza.

È ragionevole che un modello di probabilità temporale contenga un modello delle transizioni che serva a decrivere la sua evoluzione e un modello sensoriale che rappresenta i processi di osservazione realizzabili.

J43 j.04 I principali compiti dell'inferenza nei modelli temporali sono il filtraggio (per la stima dello stato), la predizione, lo smoothing o regolarizzazione e il calcolo della spiegazione più probabile. Per l'esecuzione di tutti questi compiti si conoscono semplici algoritmi ricorsivi che richiedono un tempo lineare nella lunghezza della sequenza.

Le famiglie di modelli più studiate comprendono i modelli di Markov nascosti, i filti di Kalman e le reti bayesiane dinamiche (che comprendono i due precedenti modelli come casi particolari).

L'inferenza esatta con molte variabili, quando non fanno ipotesi limitatrici come i filtri di Kalman, è computazionalmente intrattabile. Nella pratica l'algoritmo di particle filtering e i suoi derivati costituiscono una famiglia efficace di algoritmi di approssimazione.

J43 j.05 Si pone il problema di definire un linguaggio formale espressivo per i modelli probabilistici; questo problema ha interessato molti pensatori del passato: Leibniz, Jacob Bernoulli, De Morgan, Boole, Peirce, Carnap, Keynes, ... Solo recentemente sono stati studiati modelli relazionali di probabilità (RPM) e sono stati definiti e sperimentati quelli che sono chiamati linguaggi di programmazione probabilistici (PPL).

I modelli relazionali di probabilità riguardano mondi derivati dalla semantica delle basi dati per linguaggi del primo ordine e risultano adeguati quando gli oggetti dei mondi e le loro identità sono noti con certezza.

Per un modello relazionale di probabilità gli oggetti di un mondo preso in considerazione sono rappresentati da simboli di costante e le variabili casuali di base sono tutte le possibili istanze dei simboli di predicato con gli oggetti che costituiscono i vari argomenti. Quindi l'insieme dei mondi da prendere in considerazione è finito.
I modelli RPM, relational probability model, trattano in modo molto succinto mondi con grandi quantità di oggetti e si dimostrano in grado di gestire l'incertezza relazionale.

J43 j.06 I modelli probabilisticie a universo aperto (OUPM) si fondano invece sulla semantica della logica del primo ordine e ammettono vari tipi di incertezza, come incertezza dell'identità e incertezza dell'esistenza.
I programmi generativi sono rappresentazioni di vari modelli probabilistici, tra i quali gli OUPM, mediante programmi eseguibili scritti in un PPL. Un tale programma rappresenta una distribuzione sui tracciati di esecuzione del programma stesso.
I PPL hanno il pregio della capacità espressiva universale nei confronti dei modelli di probabilità.

J43 k. Decisioni semplici, complesse e multiagente

J43 k.01 Per le attività di intelligenza artifciale la teoria della probabilità serve a descrivere le credenze che un agente dovrebbe formarsi sulla base delle sue percezioni correnti e delle sue esperienze; la teoria della utilità deve invece descrivere i desideri che ogni agente esprime.

La teoria delle decisioni prende procedimenti e spunti da entrambe le suddette teorie per esaminare quello che un agente deve fare.
In linea programmatica possiamo usare la teoria delle decisioni per costruire un sistema che prende decisioni considerando tutte le azioni possibili per scegliere quella che porta all'esito considerato più favorevole; a un tale sistema si può quindi attribuire la qualifica di agente razionale.

La teoria dell'utilità mostra che un agente le cui preferenze di fronte a diverse lotterie sono in sintonia con un insieme di semplici assiomi può essere descritto come se fosse dotato di una funzione di utilità e come se scegliesse le azioni che massimizzano la utilità che si aspetta.

J43 k.02 La teoria dell'utilità multiattributo si occupa delle utilità che dipendono da più attributi distinti degli stati.

La dominanza stocastica è una tecnica che risulta particolarmente utile per prendere decisioni non ambigue, anche in assenza di valori precisi di utilità per gli attributi.
Le reti di decisioni forniscoo un semplice formalismo per esprimere e risolvere problemi decisionali; si tratta di una estensione naturale delle reti bayesiane che, oltre ai nodi di casualità, contengono nodi di decisione e nodi di utilità.

Talvolta per risolvere un problema conviene raccoglere nuove informazioni prima di prendere una decisione. Per questo si definisce valore dell'informazione come il miglioramento atteso della utilità della decisione presa dopo aver raccolta l'informazione in causa rispetto all'utilità ottenuta in assenza dei corrispondenti dati.
Questo valore risulta particolarmente utile per guidare il processo di raccolta delle informazioni prima di prendere una decisione definitiva.

Quando, come avviene spesso,è impossibile specificare la funzione di utilità dell'essere umano in modo completo e attendibile, le macchine devono operare in condizioni di incertezza circa il vero obiettivo. Questo fa la differenza quando la macchina ha la possibilità di acquisire ulteriori informazioni sulle preferenze umane.
Si trova, con una argomentazione semplice, che l'incertezza sulle preferenze comporti che la macchina deleghi le decisini all'operatore umano che la controlla, fino ad arrivare alla decisione di lasciarsi spegnere.

J43 k.03 Dopo aver visti i problemi computazionali richiesti da decisioni singole, consideriamo i problemi di decisioni sequenziali in ambienti stocastici nei quali l'utilità per l'agente dipende da una sequenza di decisioni.
Questi problemi devono affrontare, oltre alle valutazini di utilità, anche valutazioni di incertezza e percezione e comprendono come casi particolari i problemi di ricerca e pianificazione.

Si trova anche che alcuni metodi per risolverli conducono a comportamenti appropriati per ambienti stocastici.

I problemi di decisioni sequenziali in ambienti stocastici sono detti anche processi decisionali di Markov o MDP e sono definiti da un modello di transizione che specifica gli esiti probabilistici delle azioni e da

una funzione di ricompensa che determina il vantaggio conseguito con il raggiungimento di ciascuno degli stati.
Come utilità associata a una sequenza di stati raggiunti si si assume semplicemente la somma delle ricompense che si ottengono negli stati successivamente raggiunti; ad essa in talune circostanze si può aggiungere uno sconto dipendente dal tempo impiegato.

J43 k.04 La soluzione di un MDP corrisponde a una politica che associa una decisione a ogni stato raggiungibile dall'agente. Una politica ottima massimizza l'utilità della sequenza di stati toccati durante la sua esecuzione.
L'utilità di uno stato corrisponde alla somma delle ricompense attese con l'esecuzione di una politica ottima a partire da quello stesso stato.

L'algoritmo di iterazione dei valori risolve iterativamente le equazioni che collegano l'utilità di uno stato a quella degli stati suoi vicini.
L'iterazione delle politiche alterna il calcolo delle utilità degli stati rispetto alla politica corrente con il miglioramento della politica in base alle utilità correnti.

Gli MDP parzialmente osservabili, denotati con l'aacronimo POMDP, sono molto più difficili da risolvere degli MDP. Tuttavia è possibile trasformare un POMDP in un MDP nello spazio continuo degli stati-credenza. Per questi sono stati progettati algoritmi di iterazione dei valori e di iterazione delle politiche.

Il comportamento ottimo di un POMDP include la raccolta delle informazioni in grado di ridurre le incertezze e conseguentemente di prendere decisioni migliori.
Per gli ambienti POMDP è possibile costruire agenti basati sulla teoria delle decisioni. Questi agenti utilizzano una rete di decisione dinamica per rappresentare il modello di transizione e il modello sensoriale, per aggiornare il proprio stato-credenza e per consentire di registrare sequenze di azioni che possano essere riutilizzate.

J43 k.05 In molti scenari nei quali interviene la intelligenza rtificiale si incontrano agenti che devono prendere decisioni in ambienti nei quali si trovano altri agenti. Tali ambienti sono detti sistemi multiagente e gli agenti che operano in tali ambienti si dice che affrontano un problema di pianificazionesono multiagente.

Prevedibilmente la precisa caratterizzazione di ciascuno di questi problemi e le tecniche che si trovano per risolverlo dipendono fortemente dalle relazioni che intercorrono tra i vari agenti presenti.
Si distinguono soprattutto le relazioni tra agenti che cooperano dalle relazioni tra agenti che competono.

Per i vari agenti si possono definire piani congiunti; nel caso di agenti che cooperano questi piani devono essere accompagnati da qualche forma di coordinamento secondo il quale i diversi agenti devono concordare sul piano congiunto che intendono seguire.

Per definire il comportamento razionale nelle situazioni in cui interagiscono più agenti si può ricorrere alla teoria dei giochi.
Essa svolge per le decisioni multiagenti il ruolo che la teoria delle decisioni svolge per le decisioni di un agente che opera singolarmente.
Le caratteristiche delle soluzioni nella teoria dei giochi determinano i risultati razionali di un gioco, ossia i risultati che si ottengono quando tutti gli agenti agiscono razionalmente.

J43 k.06 La teoria dei giochi non cooperativi assume che gli agenti debbano prendere le rispettive decisioni con ragionamwnti indipendenti.

L'equilibrio di Nash è il genere di soluzione più importante fornito da questa teoria. Si tratta di un profilo di stategie tale che nessun agente è incentivato a cambiare la propria strategia. Per ottenerlo esistono tecniche adatte ai giochi ripetuti e tecniche adatte ai giochi sequenziali.

La teoria dei giochi cooperativi considera scenari in cui gli agenti possono fare accordi vincolanti per formare coalizioni allo scopo di cooperare. Nei giochi cooperativi i concetti di soluzione tentano di individuare le coalizioni stabili (il nucleo) e il modo per suddivide equamente il valore che una coalizione ottiene, ossia il cosiddetto valore di Shapley.

Per alcune importanti classi di decisioni multiagenti si conoscono tecniche specifiche: il contract net protocol per l'assegnazione di compiti; le aste per allocare in modi efficienti le risorse, operazione importante soprattutto se queste sono scarse; la contrattazione per raggiungere accordi per questioni di interesse comune; le procedure di voto che consentono di aggregare le preferenze.

J43 m. Cloud computing e IoT

J43 m.00 Questa sezione si occupa di alcune tecnologie sviluppate o nate nell'ambito delle ICT che presentano un forte e ampio interesse per i sistemi dell'area AI che di esse si avvalgono e che ad esse forniscono contributi.

J43 m.01 Gli algoritmi evolutivi si possono introdurre attraverso la similitudine con gli organismi viventi considerati dal punto di vista dell'evoluzione darwiniana.
Questa segnala che nella competizione per la sopravvivenza sono avvantaggiati gli individui e le popolazioni che più e meglio sanno adattarsi all'ambiente nel quale si trovano.
Similmente gli algoritmi possono meglio evolversi, ossia suggerire versioni successive di maggiore successo, quanto più consentono di valutare la propria fitness, ossia la validità dei loro risultati ottenuti per le finalità per le quali sono stati definiti (il contesto dei problemi che devono risolvere viene visto come l'ambiente nel quale operano) e quanto più consentono di ricavare spunti per correzioni migliorative (capaci di maggiore precisione, più veloci, di maggiore portata, più versatili, ...).

J43 m.03 Il cloud computing è un paradigma per la erogazione di risorse e servizi su richiesta che si serve della grande efficienza ed efficacia degli odierni strumenti telematici, sia sul piano hardware, sia su quello del software general purpose e che sta avendo notevole successo.

Si tratta di affidare grandi quantità di dati a sistemi hardware distribuiti collegati da canali di altissima portata togliendo agli utenti di questi dati molte preoccupazioni operative delle quali si fa carico il gestore del sistema cloud. Evidentemente questo sistema deve essere in grado di padroneggiare grandi quantità di memorie, grandi potenze di calcolo e una ampia gamma di sistemi software, anche molto sofisticati e specializzati.

In sostanza si tratta di un tipo di esternalizzazione di molte manovre di gestioni dei dati e in particolare di tutti i problemi derivanti dalla periodica evoluzione degli strumenti di base hardware e software.

Il vantaggio complessivo sta nel fatto che il gestore del cloud affronta un grande ventaglio di problemi che i suoi clienti dovrebbero affrontare separatamente per risolvere problemi più circoscritti. Il gestore riesce a realizzare economie di scala se riesce a gestire un grande numero di attività.

J43 m.04 Possono essere diversi gli approcci dal cloud da parte di grandi aziende e organismi pubblici da una parte e da parte di aziende medio piccole (ma non necessariamente) dall'altra.

Queste ultime adottando il cloud computing risultano fortemente vincolate alle grandi multinazionali dell'alta tecnologia, devono essere costantemente connesse al cloud, e devono preoccuparsi della possibilità che i loro dati possano essere sfruttati contro i loro interessi e quindi contro le esigenze di riservatezza dei dati delicati e critici.

Chi si affida al cloud ha il vantaggio di risparmiare sulle risorse hardware, software e umane che richiede l'autonomia.
Questo risparmio è notevolmente elevato quando si consideri la rapida obsolescenza di molte tecnologie e di molte esigenze del mercato.

Un gestore di un sistema cloud se in grado di controllare l'evoluzione tecnologica e anche i ricorrenti cambiamenti delle esigenze del mercato, delle normative statali, delle politiche dei grandi attori della tecnologia e della finanza e delle stesse vicende internazionali, riesce a realizzare notevolissimi vantaggi.
JP Questo implica che possono gestire sistemi cloud solo compagnie di altissimo profilo. In effetti i sistemi più efficienti sembra siano quelli di Amazon, Google e Microsoft (nell'ordine).

J43 m.05 Critiche al cloud computing sono formulate dai sostenitori dei contenuti aperti e in particolare da Richard Stallman; questi ha sempre sottolineato che l'offerta di servizi a bassi prezzi tende a privare i clienti del controllo sui dati e sulle attività informatiche

J43 m.06 Il termine Internet delle cose o Internet of Things (IoT) è stato introdotto nel 1999 da Kevin Ashton per prospettare la evoluzione delle rete globale come fitta rete di connessioni comprendente una elevata quantità di dispositivi fisici in grado di trasmettere dati che riguardano eventi negli ambienti fisici e sociali i quali possono rivelarsi utilizzabili per una ampia gamma di scopi.
Una tale rete avrà (e sa avendo la possibilità e la opportunità di gestire una grande massa di dati che potranno essere organizzati con le tecniche dei big data.

Per tale gestione dovranno essere messi in campo procedimenti e metododologie sulle quali avrà grande influenza l'area AI.

Le funzioni principali alle quali verosimilmente contribuirà in misura crescente l'intelligenza artificiale sono le seguenti: la eliminazione del rumore e la individuazione di dati mancanti; la omogeneizzazione, la standardizzazione, la normalizzazione e la disponibilità in svariate conversioni dei dati; i collegamenti e le integrazioni con basi dati precedentemente certificate.

Altri vantaggi di carattere generale comprendono: la possibilità di prendere decisioni rapide e più fondate grazie alla disponibilità di dati più esaurienti, meglio organizzati e aggiornati più tempestivamente; maggiore possibilità di avvalersi di sistemi di monitoraggio sui processi che coinvolgono ampi territori e molteplici elementi valutativi.

J43 n. Data mining e OCR

J43 n.01 Il termine data mining, che si può tradurre con "escavazione di dati", richiama le attività di estrazione dalle miniere di minerali di alto valore, in genere attraverso lo scavo e la movimentazione di grandi quantità di materiali e la successiva selezione di piccole quantità di materiale che presenta caratteristiche che lo rendono molto prezioso.

In termini generici possiamo dire che all'interno della elaborazione automatica dei dati le operazioni di data mining consistono nel ricercare all'interno di grandi collezioni di dati elementi utili che possono contribuire alla formazione di conoscenze giudicate utili per successive attività a carattere valutativo, oppure che consistono nel raccogliere informazioni utili per affrontare problemi specifici.
Ancora diciamo che nell'ambito della risoluzione di un problema scopo di una manovra di data mining è l'asame di dati che riguardano oggetti o processi collegati al problema i quali si suppone contengano informazioni implicite che possono essere utili per la risoluzion stessa.
Quelli che abbiamo chiamati elementi utili, informazioni utili e informazioni implicite li chiameremo "patterns".

J43 n.02 I patterns possono essere costituiti da configurazioni o schemi in grado di rappresentare raggruppamenti di dati che tendono a presentarsi con frequenza oppure schemi in grado di rappresentare relazioni che intercorrono tra questi schemi o tra questi raggruppamenti di dati.

Uno scopo successivo del data mining è quello di individuare regole con le quali esprimere i patterns e le loro relazioni.
Tutti questi patterns devono essere individuati in modo da poterli utilizzare con relativa facilità per predisli applicare a raggruppamenti di dati simili a quelli preanalizzati o addirittura dipendenti da quelli.

Gli scenari e i problemi che costituiscono le motivazioni delle operazioni di data mining appartengono ai generi più disparati, mentre le metodologie che vengono adottate si vuole che siano il più possibile indipendenti dalle motivazioni specifiche.

J43 n.03 Tipiche situazioni riguardano la ricerca di parametri di sintesi (globali) rispetto a complessi variegati di scenari specifici (che in genere riguardano località, periodi e contesti diversi).
Questi parametri di sintesi possono servire a scopi che si possono ricondurre a supporti per le decisioni.

Un esempio è dato della cosiddetta business intelligence, attività che si rivolge a un'azienda che si pone di fronte a un possibile nuovo mercato.
I patterns ricavati dall'esame di mercati precedenti possono indicare quali impegni assumere, quali prodotti sia conveniente offrire e in quali periodi dell'anno e a quali fasce di consumatori.
Problemi simili so pongono ad una azienda che prende in esame un panorama di possibili fornitori o di possibili partners con i quali accordarsi, o di candidati dipendenti da assumere, o di consulenti da ingaggiare.

Tra gli altri esempi di interventi e di motivazioni ci limitiamo a segnalare le attività elettorali, le strategie dei produttori mediatici, le iniziative assicurative e i molti problemi di controllo e di presidio sanitario.

J43 n.04 Il data mining ha natura esplorativa e in genere nelle sue molteplici applicazioni(soprattutto nelle attività su grandi volumi di dati) non ha la possibilità di basarsi sopra un modello consolidato della massa delle informazioni che deve vagliare

Quando per i dati da vagliare si dispone di un modello tendenzialmente accettabile concernente forma, contesti strutturati e valutazioni quantitative dei dati da vagliare sono disponibili metodi e procedimenti più tradizionali che sono stati sviluppati nel settore ICT chiamato analisi dei dati.

Il data mining spesso risulta determinante quando si ottengono parametri di sintesi inaspettati, poco prevedibili.
La validazione dei parametri di sintesi può risultare impegnativa e molto sensibile; questo porta alla necessità di effettuare operazioni e campagne di data mining molto meticolose e accurate.

J43 n.05 Per lo svolgimento delle indagini di data mining si distinguono due due generi di modalità (chiamate anche modelli): le modalità di verifica e le modalità di scoperta.

Le modalità di verifica prevedono che l'utente fornisca al sistema, attraverso un suggerimento, un'ipotesi e gli chieda di verificarne la validità. In questo caso il sistema di data mining non si concentra sull'estrazione delle informazioni, ma sulla verifica di ipotesi predeterminate.

Le modalità di scoperta, invece, presuppongono che il sistema individui informazioni ipotizzate nei dati a disposizione, in modo autonomo, senza ricorrere ad alcun intervento esterno.

Per esempio dall'osservazione di dati meteorologici di una certa zona si potrebbe dedurre una regola di previsione come la seguente.
(R1): "Se (la sera ci sono le nubi e la temperatura è circa 15 gradi centigradi), allora la notte pioverà".
In effetti R1 è un'informazione presente nei dati a disposizione, ma solo implicitamente, perché non è nota e non è rilevabile direttamente prima dell'attività di data mining, mentre una volta conosciuta può essere utilizzata per produrre previsioni.

In sintesi, dunque, il data mining non produce una nuova informazione, ma rende esplicita quella presente nei dati per opera di un cosiddetto "data miner', programma che esplora il contenuto di una collezione di dati (base dati, data set o un archivio simile) con l'obiettivo di estrarre regolarità esprimibili come patterns.

Può accadere che un'attività di DM identifichi regole inutili o false o debba fare fronte alla povertà o all'inesattezza di dati.

Le regole inutili sono quelle derivate da fenomeni che, pur presentendosi contemporaneamente a quelli rispecchiati dai dati, non sono loro legati da nessi causali riconosciuti come tali.
Le regole false sono in genere la conseguenza della rilevazione di dati imprecisi o spuri; questi si trovano soprattutto quando si rilevano dati in condizioni eccezionali o anomale, ossia ignorando influenze dovute a circostanze non comuni o eccezionali.

È anche possibile che per l'obiettivo di estrarre regole l'insieme di dati disponibili per la preanalisi sia condizionato da atteggiamenti specifici non dichiarati o non tenuti in debito conto.

J43 n.06 I dati su cui opera un data miner in molti casi sono organizzati in una tabella T con m righe riguardanti esempi noti e distinguibili, e con n colonne concernenti attributi che si riscontrano in tutti i diversi esempi.
Gli elementi di T possono essere numerici oppure qualitativi e simbolici. Spesso si conviene che l'ultima colonna della T, cioè l'attributo di posto n, sia quello che si vuole far dipendere dagli altri o che sia quello che si vuole classificare.

In tal caso il problema del data mining può essere formulato come segue:

Data T, trovare una funzione classificatore (f) tale che "per molti esempi" t sia

$$T[t,n] \;=\; f\left(T[t,1],...,T[t,n-1]\right) .$$

Una volta determinata f attraverso una fase di training sopra un insieme ridotto di dati, l'implementazione della fase di esercizio può essere effettuata in un opportuno linguaggio di programmazione, operazione che attualmente non presenta difficoltà.

Nel caso in cui tutti gli elementi della tabella T siano numerici, il calcolo della funzione classificatore f viene tipicamente eseguito utilizzando i metodi della statistica; nel caso di presenza di dati qualitativi, non numerici, è possibile procedere impiegando le tecniche del machine learning, ossia dell'apprendimento automatic).

Molti problemi possono essere trattati secondo entrambi i metodi.
Sia il machine learning, sia i metodi statistici partono operando su un insieme di esempi (i dati di training) e successivamente generano il classificatore, differendo però nella sua rappresentazione.
I metodi di machine learning rappresentano il classificatore usando un modello formale vicino alla programmazione, mentre i metodi statistici ricorrono a funzioni matematiche.

Un altro metodo di machine learning per rappresentare il classificatore si serve di programmi strutturati come un albero di decisione (decision tree) organizzato mediante nodi e archi.
I nodi non terminali sono contrassegnati dagli attributi della tabella T; quelli terminali dall'attributo che si vuole predire, ossia l'ultimo attributo della tabella T.
Gli archi sono contrassegnati con i valori dell'attributo che etichetta il nodo sorgente dell'arco.

I metodi statistici richiedono che tutti gli attributi siano di tipo numerico, ma per la loro significatività vengono adottato anche in casi nei quali gli attributi non sono numerici, ricorrendo a una trasformazione di queste informazioni in quantità attraverso l'adozione di opportuni accorgimenti.

J43 n.07 In statistica accade spesso di dover cercare se esiste una relazione di dipendenza di una variabile, che viene chiamata risposta o esito, da altre variabili considerate mutuamente indipendenti, che vengono chiamate predittori, covariate o variabili esplicative. A questo scopo vengono adottati vari tipi di procedimenti che presentano diversi livelli di complessità e che costituiscono la cosiddetta analisi della reggressione.

La forma più comune è la regressione lineare nella quale si cerca una retta nel caso di una variabile indipendente e un iperpiano nel caso di più variabili esplicative, che possa rappresentare la dipendenza nel modo migliore secondo una determinata valutazione numerica. La valutazione più usuale riguarda il metodo oerdinario dei minimi quadrati che ricerca la minimizzazione della somma delle differenze fra i valori osservati e i valori corrispondenti alla retta o all'iperpiano.

La regressione lineare richiede la minimizzazione della media stimata delle suddette differenze; le regressioni quantiliche invece richiedono la minimizzazione della mediana stimata o di alcuni quantili stimati.
Questo metodo richiede calcoli molto tediosi ed è stato praticato ampiamente solo dopo la disponibilità di computers di adeguata potenza.
È stato utilizzato in particolare in ecologia, campo nel quale le interazioni tra i diversi fattori sono particolarmente complesse, al fine di scoprire relazioni di valore predittivo per grandezze covariate che dipendono dai predittori in modi diversi nei diversi campi di variabilità dei predittori.

Recentemente si è imposto l'approccio noto come NCA, sigla che sta per necessary condition analysis Nella analisi dei dati denota un approccio e una tecnica recenti per la identificazione delle condizioni necessarie (ma non sufficienti) che si riscontrano in un data set. Spesso questa tecnica viene usata come complemento delle tecniche della usuale regressione, o anche delle tecniche QCA. La NCA riesce a portare chiarimenti rilevanti per la pratica e si è dimostrato che riesce a combinare rigore e rilevanza.

J43 n.08 Utilizzo delle reti neurali e training della rete

J43 n.09 Un'altro genere di applicazione del data mining chiamato "process mining" afferisce al cosiddetto "process management" e riguarda l'analisi dei processi manageriali di una azienda a partire da registrazioni di eventi di business.
Mediante opportuni algoritmi di data mining, dalle suddette registrazioni si estraggono schemi di sequenze di eventi il cui studio può contribuire a migliorare il sistema informativo della azienda interessata.

J43 n.10 Occupiamoci ora di un genere di attività che gioca un ruolo importante nella raccolta dei dati e nella organizzazione delle conoscenze e in particolare nel data mining.

L'optical character recognition, in sigla OCR, ossia il riconoscimento ottico dei caratteri riguarda la conversione dei caratteri scritti a mano o prodotti con qualche procedimento tipografico in un testo digitale gestibile da automatismi, ad esempio in un testo in caratteri ASCII o in caratteri Unicode.

I testi da trasformare in files leggibili possono essere ottenuti da molteplici tipi di documenti materiali: documenti stampati (libri, giornali, manifesti, ...), decumenti cartacei sottoposti a scannerizzazione, immagini fotografiche, didascalie di fotografie, registrazioni da schermate televisive fornite da stazioni di broadcasting o da apparati televisivi a circuito chiuso (CCTV), manoscritti di ogni genere (intimo da diari e carteggi, finanziario, amministrativo, sanitario da cartelle cliniche, da ricette, ...).
¶Questa varietà dipende dal fatto che la disponibilità di dati digitalizzati apre la possibilità di elaborazioni con ampie aspettative di applicative: riedizioni elettroniche di testi, ricerche con i più ampi obiettivi (linguistici, sociologici, storici, computazioni cognitive, traduzioni automatiche, trasformazione da testo a parlato).

Questo fa dell'OCR un campo nel quale convergono ricerche di pattern recognition, di visione artificiale, di lessicografiq e di machine learning.

J43 n.11 Le prime ricerche del settore si collegano ad attività telegrafiche e alla ricerca di dispositivi di lettura per ciechi e per ipovedenti.
Successivi progressi hanno portato verso il 1930 ad applicazioni concernenti la gestione e l'analisi dei supporti di dati allora i più capienti ed effcienti, i microfilms.

Negli anni 1970 si sono messi a punto strumenti sensibilmente più efficienti e di diffusione rilevante ad opera di vari ricercatori e in particolare di Ray Kurzweil. Al successo dei relativi prodotti hanno contribuito anche lo sviluppo di dispositivi per la scannerizzazione chiamati "flat bed scanners" e la costruzione di sintetizzatori per la traduzione da testo a parlato.

Nel 2000 sono stati realizzati servizi in linea (WebOCR), ottenibili da ambienti di cloud computing; intorno al 2010 si sono resi disponibili strumenti molto efficienti che si servono di smartphones e di smartglasses.

Inoltre si possono inserire prestazioni di OCR in molti prodotti ottenuti con sistemi di sviluppo software servendosi delle interfacce chiamate OCR API.

Molti sistemi di scrittura possono avere in dotazione sistemi OCR, sia proprietari che oper source.

Sono ormai numerose Le lingue trattabili con strumenti OCR:nei caratteri latini, cirillici, arabi, ebraici, bengali, devanagari, tamil, cinesi, giapponesi e coreani.

Occorre aggiungere che l'accuratezza dei procedimenti vede notevoli differenze.
Mentre per i testi stampati in lingue che usano caratteri latini è piuttosto elevata, si abbassa quando si tratta di lingue con caratteri dall'aspetto sensibilmente più elaborato.

Vi sono anche sistemi OCR in grado di riconoscere la scrittura a mano, attività più impegativa in quanto la scrittura umana può essere molto disordinata e disomogenea.
In questi campi sono necessarie nuove attività di ricerca.

J43 o. Sistemi esperti

J43 o.01 Si tratta di programmi piuttosto elaborati che hanno per fine la risoluzione di problemi complessi che tradizionalmente richiedono l'intervento di persone esperte del settore al quale afferiscono i problemi suddetti.

Ciascuno di questi programmi "esperti" è in grado di avvalersi di meccanismi inferenziali che consentono di ricavare dalle richieste di chi pone il problema gli elementi di partenza per i procedimenri risolutivi che il programma esperto sa mettere in campo.

In un sistema esperto si riconoscono tre componenti operative.

(1) Una base di conoscenza nella quale si trovano le nozioni delle quli si può servire, le regole deduttive che può utilizzare e le procedure delle quali si può servire.

(2) Un cosiddetto motore inferenziale che implementa gli algoritmi deduttivi che consentono di effettuare le trasformazioni necessarie al suo lavoro di ricerca di soluzioni.

(3) Una interfaccia per i suoi utenti interattivi, le persone interessate alle istanze dei problemi trattabili che possono influenzare le sue manovre (o circoscrivendole, o ampliandole, oppure riorientandole).

Può accadere che le ricerche di una soluzione richedano di esaminare gamme di possibilità estremamente ampie, tali da richiedere tempi di indagine inaccettabilmente lunghi.
In queste situazioni molti sistemi esperti assumono comportamenti euristici spesso sviluppando deduzioni probabilistiche mediante logiche fuzzy, rinunciando a ottenere soluzioni ottimali per giungere a risultati chiamati subottimali, caratterizzati da alte probabilità ma che rischiano la fallibilità e che auspicabilmente risultano utili a proseguire nella soluzione del problema mediante altri procedimenti.

J43 o.02 Un primo grande gruppo di sistemi esperti si basa su regole aventi la forma

`if` *clausola condizionale* `then` *azione primaria* `else` *azione alternativa*

.

Con tali regole si possono controllare interventi su tipologie di situazioni diagnosticabili abbastanza agevolmente e anche ampie e articolate; in particolare in tal modo vengono affrontati interventi terapeutici.

Sono disponibili sistemi di questo genere semplici destinati a scopi didattici e sistemi progettati per la generazione di sistemi esperti specialistici.

J43 o.03 Un secondo gruppo è costituito dai sistemi esperti he si servono di alberi.

Uno di tali sistemi applica a un insieme di dati dei meccanismi selettivi e deduttivi per ricavare un albero di classificazione dei dati stessi.

È opportuno osservare che un sistema esperto non va considerato un programma creativo che giunge a proporre qualcosa di innovativo per la soluzione del problema che gli viene posto.
Esso, più modestamente, fa riferimento a un operatore umano esperto del settore e presenta il vantaggio di riuscire a elaborare grandi quantità di dati, di classificarli secondo una vasta scelta di criteri e di prendere in considerazione moltissimi dettagli, senza i rischi di distrazioni, di momenti di fatica e di conseguenti imprecisioni che presenta ogni agente umano.

J43 o.04 Vediamo un elenco sintetico delle modalità operative dei sistemi esperti realizzati.

Attività predittiva: deduzione delle possibili situazioni che si possono raggiungere da un determinato elenco di dati iniziali.

Diagnosi: individuare i possibili malfunzionamenti di un sistema industriale o organizzativo conseguenti dalla osservazione dei livelli di funzionamento ricavabili da un opportuno gruppo di punti di osservazione del sistema stesso.

Progettazione: determinare una migliore configurazione di oggetti da porre in mutua relazione, rispettando dei vincoli e tenendo conto di più parametri valutativi.

Pianificazione: determinazione di azioni in grado di modificare uno scenario in una direzione voluta.

Monitorggio: effettuare osservazioni su un ambiente vulnerabile, segnalare le possibili falle e organizzare manovre di intervento preventivo e difensivo.

Debugging: proporre interventi incrementali per evitare situazioni problematiche, adpoperandosi per ottenere miglioramenti progessivi.

Repristino: segnalazione di comportamenti difettosi di un sistema e decisione di interventi di riparazione.

Controllo: predizione di situazioni critiche di un sistema oggetto di monitoraggio e decisione di interventi di riequilibrio.

Istruzione: valutazione dei risultati di studenti, classificazione dei loro comportamenti e interventi didattici correttivi o migliorativi.

J43 p. Reti neurali

J43 p.01 In genere si dovrebbe distinguere tra le reti neurali biologiche facenti parte dei sistemi nervosi degli umani e di altri animali e le reti neurali artificiali costruite con materiali forniti dalla elettronica e da comparti industriali contigui che dalle biologiche sono state ispirate.
In queste pagine useremo spesso il termine "rete neurale" come abbreviazione di rete neurale artificiale e non parleremo molto delle reti neurali biologiche.
Segnaliamo anche che nella lingua inglese e nella letteratura scientifica si parla di artificial neural network, termine abbreviato con l'acronimo ANN.

Le reti neurali sono modelli computazionali caratterizzati da una rete di unità semplici, i nodi della rete (chiamati anche stati, o neuroni) i quali sono collegati da connessioni definite in modo da poter rappresentare le loro mutue influenze.

Si distinguono nodi di ingresso, nodi interni e nodi di uscita e una rete neurale si può pensare come uno strumento che a funzioni realizzate da tensioni elettriche variabili nel tempo applicate ai nodi di ingresso facciano corrispondere funzioni analoghe fornite dai vari nodi di uscita, in conseguenza delle variazioni delle tensioni che si riscontrano nei nodi interni e che sono determinate dal complesso delle interconnessioni.
La corrispondenza tra tensioni in entrata e tensioni in uscita fa di ogni rete neurale un modello matematico di uno strumento di calcolo.

Le connessioni tra i nodi di una ANN nel modo più semplice sono schematizzate da un digrafo, cioè da un insieme di nodi e da un insieme di coppie di nodi esprimenti collegamenti tra nodi e chiamate archi.

Il digrafo di una rete può presentare o meno dei cicli; le reti con cicli sono dette "reti neurali ricorrenti".

J43 p.02 Le reti neurali serono a determinare funzioni che approssimano funzioni complesse $\mathbf{f}$ del genere che ad una m-upla di valori reali $\mathbf{x}$ associano una n-upla di valori reali $\mathbf{y}$ quando si conoscono solo poche coppie della forma $\langle \mathbf{x}, f(\mathbf{x}) \rangle$ oppure si conoscono solo le variabili $\mathbf{x}$ e i generi dei valori $\mathbf{y}$ in quanto adatti a essere visti come cause di conseguenze che si intendono spiegare in termini di valori iniziali dei tipi dei componenti degli $\mathbf{x}$.

Queste funzioni f si cercano fra quelle esprimibili come composizioni di variabili reali intermedie g_1, g_2, ..., g_k che a loro volta si possano esprimere ome composizoni di altre variabili intermedie h_1, h_2, ...h_j e così via per un certo numero s di "strati" successivi.

Questa situazione si può schematizzare con un digrafo che vede nodi più a sinistra che rappresentano le componenti di $\mathbf{x}$, nodi più a destra che rappresentano $\mathbf{f}(\mathbf{x})$ e le funzioni intermedie ripartite per strati successivi da sinistra a destra finendo con lo strato delle h_j, più a sinistra dello strato delle g_k che precede lo strato della $\mathbf{f}$.

Una prima semplificazione impone che le funzioni di uno strato possono essere ottenute solo da composizioni di funzioni dello strato precedente, quello immediatamente più a sinistra.
Questa situazione si schematizza con un digrafo i cui nodi si dispongono nei successivi strati e con archi che possono andare solo da uno strato all'immediatamente successivo e tali che gli archi che entrano in un nodo dicono quali funzioni alla sua sinistra entrano nella composizione che lo eprime.
Situazioni più complesse consentono archi che non rispettano la regola precedente, archi che saltano qualche strato, che vanno verso sinistra e che quindi consentono la formazione di cicli (e anche di cappi).

J43 p.03 Queste situazioni possono essere simulate da circuiti elettrici nei quali i valori variabili dei nodi sono rappresentati da tensioni elettriche variabili e gli archi da collegamenti elettrici che possono essere attenuati; inoltre le correnti che entrano in un nodo possono influire diversamente ovvero essere caratterizzate da pesi relativi.

Questo tipo di simulazione è stato suggerito dalla rete dei neuroni biologicici ciascuno dei quali presenta un soma centrale contenente il nucleo al quale sono collegati i dendriti e l'assone, un ottimo onduttore elettrico che presenta numerose diramazioni che conducono ai neuroni (anche migliaia) che il nostro influenza.
I neuroni influenzati ricevono il segnale dell'influenzante attraverso i loro dendriti che ricevono il segnale già attenuato e lo inviano al proprio nucleo ulteriormente ridotto.
Con questo meccanismo di invio di segnali da un neurone ad un altro si generano trasmissioni, circuiti e influenze complesse che costuituiscono la estremamente complessa attività cerebrale.

L'algoritmo con il quale si simula come viene influenzato un neurone deve deve accogliere tutti i valori di input x_i con i loro pesi relativi, che denotiamo con w_i, deve comporli e deve tenere conto di una funzione di attivazione che denotiamo con $\alpha(x)$ per tenere conto della funzione di sparo dei neuroni influenzanti.

Una tipica espressione utilizzata si serve solo delle operazioni somma e prodotto e assume una forma del tipo

$$y(x) \;=\; \alpha(x) \left(\sum_{j=1}^{m} w_j \, x_j \right) \;.$$

Tipiche funzioni di attuazione sono la unzione a scalino, la tangente iperbolica, la funzioe logistica, la funzione ReLU (rectified linear unit) e l funzione softmax.

J43 p.04 Questi modelli matematici sono utilizzati per tentare di risolvere problemi che si presentano in diversi ambiti tecnologici: in elettronica, per computazioni digitali, per vari tipi di simulazioni e in particolare per simulare reti neurali biologiche.

Molti studiosi sostengono l'opportunità di affrontare la gran parte dei problemi impegnativi servendosi delle reti neurali, ossia giocando sulle caratteristiche trasformazionali che si possono ottenere da questo tipo di circuiti.

Le reti neurali si possono realizzare sia mediante programmi software per i vari tipi di computers, sia mediante dispositivi hardware dedicati presi in considerazione nell'ambito del Digital Signal Processing. Spesso le reti neurali esse vengono utilizzate con metodi computazionali della logica fuzzy.

Ricordiamo che i nodi componenti elementari delle reti neurali, chiamati anche neuroni artificiali, sono stati proposti fin dal 1943 da W.S. McCulloch e Walter Pitts.

Le reti neurali vengono utilizzate come strumenti adattabili e per adattarle vengono loro applicate varie modalità di addestramento; in corrispondenza di esse si parla di modalità di apprendimento delle reti.

Per queste modalità si seguono tre paradigmi: apprendimento supervisionato, apprendimento non supervisionato e apprendimento per rinforzo.

J43 p.05 L'apprendimento supervisionato richiede la disponibilità di dati per l'addestramento costituiti da esempi significativi di coppie $\langle$ dati di ingresso, risultati in uscita $\rangle$. Trattando questi esempi la rete può imparare ad inferire la relazione che collega i primi membri ai secondi.

Successivamente, la rete è addestrata mediante un opportuno algoritmo (tipicamente, la backpropagation, che è appunto un algoritmo d apprendimento supervisionato), il quale usa tali dati allo scopo di modificare i pesi e altri parametri della rete stessa in modo tale Se l'addestramento ha successo, la rete impara a riconoscere la relazione incognita che lega le variabili di ingresso a quelle di uscita, ed è quindi in grado di fare previsioni anche laddove l'uscita non è nota a priori; in altri termini, l'obiettivo finale dell'apprendimento supervisionato è la previsione del valore dell'uscita per ogni valore valido dell'ingresso, basandosi soltanto su un numero limitato di esempi di corrispondenza (vale a dire, coppie di valori input-output). Per fare ciò, la rete deve essere infine dotata di un'adeguata capacità di generalizzazione, con riferimento a casi ad essa ignoti. Ciò consente di risolvere problemi di regressione o classificazione.

J43 p.06 Un apprendimento non supervisionato basato su algoritmi di addestramento che modificano i pesi della rete facendo esclusivamente riferimento ad un insieme di dati che include le sole variabili di ingresso. Tali algoritmi tentano di raggruppare i dati di ingresso e di individuare pertanto degli opportuni cluster rappresentativi dei dati stessi, facendo uso tipicamente di metodi topologici o probabilistici. L'apprendimento non supervisionato è anche impiegato per sviluppare tecniche di compressione dei dati.

J43 p.07 L'apprendimento per rinforzo nel quale un opportuno algoritmo si prefigge lo scopo di individuare un certo modus operandi, a partire da un processo di osservazione dell'ambiente esterno; ogni azione ha un impatto sull'ambiente, e l'ambiente produce una retroazione che guida l'algoritmo stesso nel processo di apprendimento. Tale classe di problemi postula un agente, dotato di capacità di percezione, che esplora un ambiente nel quale intraprende una serie di azioni. L'ambiente stesso fornisce in risposta un incentivo o un disincentivo, secondo i casi. Gli algoritmi per il reinforcement learning tentano in definitiva di determinare una politica tesa a massimizzare gli incentivi cumulati ricevuti dall'agente nel corso della sua esplorazione del problema. L'apprendimento con rinforzo differisce da quello supervisionato poiché non sono mai presentate delle coppie input-output di esempi noti, né si procede alla correzione esplicita di azioni subottimali. Inoltre, l'algoritmo è focalizzato sulla prestazione in linea, la quale implica un bilanciamento tra esplorazione di situazioni ignote e sfruttamento della conoscenza corrente.

J43 q. Elaborazioni numeriche, grafiche, algebriche e logiche

J43 q.01 I computers sono stati progettati e utilizzati con lo scopo di eseguire calcoli numerici, negli anni 1940 per scopi bellici, per poche attività produttive e per pochi fini scientifici.
Negli anni 1950 sono serviti principalmente per effettuare elaborazioni di interesse amministrativo, soprattutto per effettuare la contabilità aziendale.

Intorno al 1960 sono stati fatti i primi tentativi di calcoli algebrici e simbolici (linguaggi Schoonschip per calcoli di fisica delle particelle e Formac).
Nel 1964 è stato definito da Carl Engelman MATHLAB, sistema computazionale munito di alcuni dei primi strumenti dell'intelligenza artificiale.

I primi sistemi con prestazioni di calcolo simbolico che hanno conseguito popolarità sono stati muMATH, Reduce, Derive e Macsima.
Alcuni di questi sono successivamente passati all'area del free software.
Essi hanno contribuito ad ampliare le tavole per i calcoli simbolici, riguardanti sorattutto funzioni speciali, equazioni differenziali e calcolo di integrali; va segnalato che questi sistemi hanno consentito di rendere le tavole tradizionali più precise eliminando errori difficili da individuare anche dagli specialisti.

J43 q.02 Per vari anni i sistemi più di successo comerciale sono stati Mathematica e Maple. Si tratta di sistemi con ampie gamme di prestazioni numeriche, simboliche, grafiche e di animazione. Ad essi si affianca il sistema SageMath liberamente disponibile.

Altri sistemi sono specializzati in aree più circoscritte. Tra questi ricordiamo CoCoA rivolto all'alghebra commutativa, FORM per calcoli sulle particelle elementari, GAP rivolto alla teoria dei gruppi e alla combinatorica, Macauly2 per la geometria algebrica e l'algebra commutativa, Magma nato per la teoria dei gruppi che ha poi ampliato i suoi obiettivi e Wolfram Alpha utilizzabile online.

Nel settore del calcolo numerico ha ancora ampio utilizzo MATLAB (sta per matrix laboratory) munito di possibilità grafiche curate soprattutto per la produzione di grafici che possono essere facilmente inseriti in rapporti tecnici e in serie di slides per attività di presentazione ai livelli didattico, specialistico e industriale.

Recentemente i sistemi di computer algebra sono stati implementati servendosi di reti neurali artificiali.

A partire da 1987 sono stati dotati di funzioni CAS varie calcolatrici tascabili di largo uso (Hewlet Packard, Texas, Casio, ...).

Tra i sistemi di calcolo per fini relativamente circoscritti è notevole il successo di "R" negli ambienti che devono affrontare problemi di statistica nei molteplici campi applicativi di questa disciplina.

J43 q.03 Diamo un eleco sintetico delle tipiche prestazioni di manipolazione simbolica ottenibili dai CAS.

Semplificazione delle espressioni matematiche, anche tenendo conto di vincoli.

Sostituzione di simboli e particolarizzazione di simboli variabili con valori numerici.

Operazioni di fattorizzazione di polinomi e di trasformazioni riguardanti frazioni parziali.

Soluzioni di equazioni differenziali e integro-differenziali.

Editing di espressioni matematica bidimensionali.

Plotting di diagrammi bi e tridimensionali con possibili animazioni.

Plotting di mappe e di diagrammi schematici.

J43 q.04 Generazione di codice ottimizzato per routines di calcolo.

Possibilità di controllare operazioni aritmetiche in precisione arbitraria e calcoli su numeri interi molto elevati riguardanti in particolare operazioni dettate dalla teoria dei numeri.

Possibilità di servirsi di un linguaggio di programmazione che consente agli utenti di implementare propri algoritmi.

Possibilità di interfacciare librerie di routines numeriche e statistiche.

Possibili aggiunte da utilizzare in settori specialistici della matematica applicata, in particolare per la fisica teorica, la chimica computazionale e la bioinformatica.

J43 q.05 Attualmente i sistemi di calcolo numerico, grafico e simbolico costituiscono strumenti che forniscono risultati molto affidabili, che sono ben conosciuti da numerosi professionisti e che possono essere interfacciati con molti sistemi applicativi servendosi di tecniche dell'ingegreria del software altrettanto ben consolidate.

Molti sistemi sono dotati di API, application programming interfaces, che rendono agevole la collazioni di sottosistemi software per ottenere sistemi di calcolo altamente produttivi.

Gran parte degli studi di ingegneria, di architettura e delle scienze applicate fanno uso corrente dei sistemi sopra segnalati.

Occorre tuttavia avvertire che in molte applicazioni si possono incontrare problemi molto impegnativi che promettono di ottenere risultati di elevato valore, ma richiedono ulteriori studi specialistici e ricerche innovative che potrebbero risultare costose e non garantire risultati in tempi prevedibili.

J43 q.06 Vengono sviluppati sistemi software per la dimostrazione di nuovi teoremi che si servono di elaborazioni logiche che possono essere molto impegnative e imprevedibili nelle conclusioni.

Sono disponibili anche sistemi per la verifica di teoremi.
Questi strumenti vengono utilizzati in molti settori della matematica sperimentale.

J43 q.07 Ricordiamo anche alcune altre prestazioni dei sistemi CAS.

Manipolazione di stringhe per fini di matching e ricerca nei testi e in basi dati di scritti e traduzioni testuali di registrazioni di parlato in linguaggio corrente.

Produzione e editing di immagini generate da computer ed elaborazione di segnali mediante manipolazione di immagini.

Sintesi sonore.

Tutte queste attività portano il settore del calcolo simbolico, numerico e grafico a mantenere forti collegamenti con vari settori della ricerca nella intelligenza artificiale.

J43 q.08 Viene organizzata annualmente una Conference on Automated Deduction (CADE) nella quale si confrontano le efficacie e le efficienze di diversi dimostratori di teoremi.

Tra questi segnaliamo Prover9, piuttosto diffuso soprattutto grazie alla sua capacità doi presentare le dimostrazioni trovate in forme abbastanza leggibili.

J43 r. Elaborazioni del linguaggio naturale e chatbots

J43 r.01 Il controllo dei linguaggi naturali mediante computers serve agli agenti e ai sistemi AI per tre scopi principali.
(1) Consentire agli agenti e ai sistemi di comunicare con gli umani attraverso scritti e parlato, i mezzi di comunicazione utilizzati dalle persone più facilmente e più abitualmente.
(2) Apprendere, ossia aumentare le proprie basi di conoscenze a partire dalla grande mole di informazioni registrate in linguaggio naturale (Wikipedia conta più di 30 milioni di pagine), molto più estese delle conoscenze accumulate attraverso modi di esprimersi artificiali (contenuti di archivi strutturati, formule matematice e della logica formale).
(3) Far procedere la capacità di comprendere situazioni complesse e la sua utilizzazione attraverso automatismi e in particolare attraverso strumenti afferenti alla AI dei linguaggi naturali, in sintonia con gli studi della linguistica, della psicologia cognitiva e delle neuroscienze.

In una prospettiva di più lungo termine si pensa alla crescita di una documentazione affidabile, ampiamente fruibile e utile al miglioramento della vita delle persone che si auspica possa venire ampiamente giudicata elemento essenziale per la crescita della società nel suo complesso.

In particolare si pensa alla documentazione scientifica e tecnologica, già dotata di un alto livello di verificabilità a falsificabilità e quindi altamente affidabile, nonché riutilizzabile efficientemente.

J43 r.02 Ogni linguaggio naturale costituisce un campo fenomenico assai complesso e il controllo di più linguaggi naturali è sempre stato un problema estremamente impegnativo.

Per tenere sotto controllo un linguaggio occorre esaminarlo da quattro punti di vista principali: lessicale, sintattico, semantico e pragmatico; questo vale soprattutto se lo si controlla per obiettivi pratici di ampia portata e in particolare per attività dell'area AI.

Dal punto di vista lessicale una espressioe del linguaggio, ad esempio una frase, è costituita da una sequenza di stringhe, le parole, sopra un alfabeto che in genere è definito in modo accurato. Accade anche che linguaggi diversi in genere richiedono alfabeti diversi che talora differiscono per pochi segni, ma che possono anche presentare grandi differenze.

Il punto di vista sintattico attribuisce ad ogni frase una struttura formale: questa è ben definita per il linguaggio delle espressioni aritmetiche riguardante numeri e operandi espliciti o indeterminati, ma è sensibilmente più elaborata per le espressioni matematiche più generali e ancor più impegnativa per le frasi dei linguaggi naturali di largo uso.

Il punto di vista semantico prevede che ogni parola e ogni frase sia portatrice di significato. Il collegamento tra parole e frasi e significati risulta molto impegnativo, sia per l'ampiezza e la varietà dei possibili significati, sia per le molte possibili indeterminatezze e ambiguità.

Alcune delle vaghezze semantiche e delle ambiguità sono contrastabili (eliminabili o riducibili) tenendo conto del contesto, altre osservando la struttura sintattica, ma molti casi costituiscono delle difficoltà anche gravi.

Molte ambiguità possono provenire dal parlante umano. Possono dipendere dalle diverse formazioni dei parlanti, dalle loro intenzioni correnti e dagli ambienti nei quali si esprimono.
Spesso i parlanti umani sono vaghi nella scelta dei termini, trascurano elementi determinanti per chiarire il contesto, si servono di metafore che possono risultare soggettive o presumono sottintesi.

Il punto di vista pragmatico coinvolge le finalità delle espressioni e dei loro contesti comunicativi. Su quanto pronunciato o scritto influiscono molto gli obiettivi del parlante, mentre sulle analisi che gli

aspiranti controllori effettuano sulle espressioni e i testi da esaminare possono essere determinanti gli scopi che si prefiggono i controllori stessi.

J43 r.03 L'analisi dei linguaggi naturali per scopi di AI, necessariamente, procede attraverso modelli nei quali, a causa della complessità della tematica, non possono mancare elementi probabilistici.

Il controllo di un linguaggio in genere fa riferimento a un complesso di distribuzioni di probabilità che alle singole parole o alle brevi sequenze di parole attribuiscono valutazioni probabilistiche della loro rispondenza e funzionalità rispetto a requisiti spesso difficili da decifrare.
Un esempio è dato da espressioni della forma $\mathcal{P}(tematica, w)$ che attribuiscono una probabilità che la parola w appartenga al campo tematico specificato che potrebbe riguardare settori molto diversi: affari, sociologia, gossip, etica, estetica, meteo, ...
La somma sulle tematiche di queste probabilità assegnate a ogni parola w deve essere pari a 1.

Spesso risulta utile il cosiddetto modello Bayes ingenuo; questo per una frase di n parole verte sopra un'espressione della forma

$\mathcal{P}(tematica, w_1 w_2 ... w_n) \;=\; \alpha \mathcal{P}(tematica) / cdot \prod_{i=1}^{n} \mathcal{P}(tematica, w_i)$,

dove $\mathcal{P}(tematica)$ esprime una valutazione del peso probabilistico di ogni tematica che potrebbe essere richiamata dal parlante.

Un tipo di approccio alla comprensione del linguaggio naturale si serve delle cosiddette borse di parole (bags). Per ciascuna delle tematiche che possono intervenire si costruisce una cosiddetta borsa di parole contenente parole che caratterizzano la tematica stessa, ciascuna con un peso esprimente la frequenza del suo uso.
Queste borse vengono ottenute da un corpus testuale di addestramento nel quale ogni frase è attribuita a una tematica.

Spesso si usano corpora di milioni di occorrenze di parole, ma si sono organizzati anche addestramenti con 2.5 miliardi di occorrenze di parole tratte da Wikipedia e con 14 miliardi di occorrenze di parole ricavate da pagine Web di diversi generi.
Da un corpus si ricavano le probabilità espresse da $\mathcal{P}(tematica)$.

Occorre osservare che i risultati di questi addestramenti sono onerosi: si potrebbero richiedere distribuzioni di probabilità per le occorrenze di 100 000 parole per le varie tematiche.
Anche per questi risultati è opportuno trascurare sia parole molto frequenti che parole molto rare.

J43 r.04 Le tecniche di borsa di parole non tengono conto delle correlazioni tra le parole che si susseguono nelle frasi.
Per tenerne conto viene adottato un tipo di tecnica basato su catene di Markov che tiene conto della dipendenza tra parole successive nelle frasi.

Si distinguono i metodi a k-grammi nei quali si tiene conto della dipendenza di una parola dalle k parole che la precedono con $k = 2, 3, ...$ e per questo si usano espressioni della forma

$\mathcal{P}(w_j | w_1 ... w_{j-1}) \;=\; \mathcal{P}(w_{j-k+1} ... w_{j-1}) \quad \text{e} \quad \mathcal{P}(w_1 ... w_n) \;=\; \prod_{j=2}^{n} \mathcal{P}(w_j | w_{j-k+1} ... w_{j-1})$

Questi metodi hanno successo per obiettivi pratici come il rilevamento dello spam, la classificazione di recensioni di prodotti come positive o negative e le attribuzioni di brani ad autori non esplicitati attraverso la individuazione degli stili personali.

Metodi analoghi riguardano l'analisi del succedersi dei caratteri nelle parole, studio svolto per riuscire a identificare i linguaggi cui appartengono le parole di testi registrati per scopi che riguardano indagini che possono essere giudiziarie, sicuritarie, filologiche o archeologiche.

Per raggiungere questi obiettivi possono bastare testi ridotti a poche parole.

In alcuni casi risultano efficaci i metodi skip-gram, saltagrammi, nei quali si correlano parole non adiacenti ma separate solo da un'altra parola o da poche parole.

J43 r.05 Per le parole che risultano sconosciute dopo un addestramento, per evitare che d esse siano assegnate probabilità nulle che non consentirebbero di migliorare la comprensione, risulta opportuno trattarle tutte come se fossero una unica parola identificata da un apposito simbolo, ad esempio $\langle UNK \rangle$.

Talvolta risulta più conveniente usare diversi simboli sostitutivi come $\langle NUM \rangle$ per le scritture numeriche e $\langle EMAIL \rangle$ per gli indirizzi di posta elettronica.
Inoltre si rivela spesso utile il simbolo $\langle S \rangle$ per indicare inizio e fine frase per poterli trattare, risp., come w_0 e w_{n+1}.

Altri tipi di accorgimenti riguardano la regolarizzazione o smoothing, che consiste nel trattare come equivalenti i k-grammi che presentano determinate somiglianze, e il metodo di back-off o retrocessione che, quando si incontrano k-grammi poco frequenti o assenti dall'addestramento, retrocede a trattare i $k-1$-grammi.
Viene anche adottata la cosiddetta regolarizzazione con interpolazione lineare che si rivolge ai trigrammi poco frequenti aumentando la loro probabilità con un espressione lineare che coinvolge anche i loro suffissi della forma

$$\mathcal{P}(c_i|c_{i-2:i-1}) \;\rightarrow\; \lambda_3\,\mathcal{P}(c_i|c_{i-2:i-1}) + \lambda_2\,\mathcal{P}(c_i|c_{i-1}) + \lambda_1\,\mathcal{P}(c_i)\ ,$$

con i coefficienti tali che sia $\lambda_3 + \lambda_2 + \lambda_1 = 1$ e selezionati in modo da massimizzare l'aspettativa o con la ricerca automatica o tenendo conto della frequenza constatata per il trigramma $c_{i-2:i}$.

J43 r.05 Mostriamoo alcuni semplici esempi.

Consideriamo le due espressioni "un gatto nero" e "gatto nero un". Un sistema che fa riferimento a dati di addestramento sceglie la prima perché la seconda segue uno schema che non si incontra mai. Un parlante nativo sceglie la prima perché segue lo schema familiare articolo-aggettivo-sostantivo, contrariamente alla seconda.

Consideriamo poi l'espressione "il gattino giocoso"; anch'essa risulta familiare a chi ha una buona pratica dell'italiano. Un non italofono che non ha mai incontrato giocoso può essere portata a riconoscere che si tratta di un aggettivo come tanti ottenuti da una parola nota e dalla desinenza "oso": affettuoso", rancoroso", "impetuoso", ...
Inoltre la prossimità sintattica tra "un" e "il" avvicina l'ultima espressione alla "un gatto nero" che potrebbe esserestata già evidenziata tra dati di addestramento.

Occorre osservare che i metodi basati sulla frequenza delle parole scontano il fatto che le parole sono trattate come atomi, privi di distinzioni interne.
Per questo si possono adottare modelli che trattano parole strutturate o parole fattorizzate. Tra questi il modello detto "word embedding" fa riferimento a un vocabolario che a ogni parola riconosciuta associa più significati ed eventualmente più categorie sintattiche aprendo la possibilità di procedimenti più articolati e di portata più estesa e più incisiva.

J43 r.06 Un metodo efficace consiste nel trattare parole associate ad una POS, part of speech (POS) o a più d'una. Per questo genere di entità si usa anche il termine "tag" e la attribuzione di una POS a una parola si chiama part of speech tagging.

Sono oggi disponibili corpora di milioni di parole dotate di POS e con una tale disponibilità si può avviare l'analisi di una frase con la part of speech tagging.
Questi tags sono qualifiche sintattiche schematizzate che possono sollevare critiche dai linguisti, ma servono per procedere in attività con una utilità potenzialmente molto rilevante come la risposta a domande e la traduzione automatica.

J43 r.07 Modello di Markov nascosto o HMM, hidden Markov model
Mediante l'algoritmo di Viterbi raggiunge accuratezza molto elevata intorno al 97%.
modelli generativi e modelli discriminativi

J43 r.08 L'elaborazione dei linguaggi naturali risulta essenziale in numerosi campi dell'area AI.
Il riconoscimento vocale consiste nel trasformare brani di parlato nei corrispondenti scritti e i sistemi attualmente disponibili sono decisamente affidabili, in quanto presentano percentuali di errore tra il 3% e il 5%.

La sintesi vocale è il processo inverso del riconoscimento. I migliori sistemi attuali riescono a pronunciare correttamente le singole parole e ottengono un flusso del parlato con giuste pause e sottolineature efficaci. Si realizzano anche documenti parlati con il contributo di più voci.

Alla traduzione automatica sono dedicati numerosi sistemi. In genere sono addestrati con un corpus bilingue costituito da coppie di documenti tendenzialmente equivalenti.

Questi sistemi sono migliorati notevolmente con l'adozione delle reti neurali ricorrenti sequenza-sequenza e con l'utilizzo del meccanismo di focalizzazione dell'attenzione riuscendo ad avvicinare, per talune coppie di linguaggi, le prestazioni dei traduttori umani professuinisti.

J43 r.09 Estrazione delle informazioni

Information retrieval

Il compito di risposta a domande
J43 r.10 Lo studio delle lingue naturali richiede di tenere ben presente il punto di vista semantico: si hanno stringhe e linguaggi dotati di significati e si pone il problema di trovare meccanismi che associno a ciascuna stringa di un linguaggio un suo significato, il quale a sua volta va espresso con termini che devono risultare precisi e maneggevoli il più possibile.
Questo spesso mon si riesce a ottenere e il significato viene espresso con frasi intuitive, suggestive, ricorrendo a metafore e ad altre figure retoriche.

Occorre considerare la tendenza odierna a rielaborare anche i significati delle frasi per vari scopi, per realizzare sistemi automatici con gamme di prestazioni sempre più estese.

Una prestazione particolarmente richiesta è la soluzione del problema della traduzione automatica.

Tutte queste esigenze si possono affrontare con notevole efficacia operando prevalentemente sul piano formale, mediante i modelli formali degli oggetti, dei processi e degli scenari di interesse applicativo.
Inoltre oggi possiamo applicare ai sistemi formali più complessi strumenti informatici con alta efficienza e versatilità: un programma per il computer si può considerare una macchina formale che trasforma una stringa in ingresso, constatato che abbia forma legale e che esprima i dati di un problema in una stringa in uscita rappresentante il risultato che risolve la corrispondente istanza di problema.

J43 r.11 In questa ottica ci si trova davanti alla necessità di studiare l'aspetto pragmatico dei linguaggi in quanto insiemi di stringhe esprimenti significati.

Questo vale anche per i linguaggi artificiali, ma risulta particolarmente impegnativo per i linguaggi naturali.

Una situazione risolvibile con relativa facilità riguarda il trattamento dell'ambiguità per le stringhe dei linguaggi generati da grammatiche a struttura di frase del genere context free, acontestuale.

Queste stringhe si possono arricchire con il corrispondente albero sintattico o equivalentemente si possono arricchire con coppie di parentesi coniugate che evidenziano le sottostringhe ottenibili con la generazione.

Consideriamo la frase esemplare:

la confutazione della affermazione dell'avversario

Questa si può interpretare in due modi contrapposti: si tratta di un soggetto implicito che confuta l'affermazione dell'avversario oppure si tratta dell'avversario che confuta l'affermazione di un soggetto implicito?

Questo dilemma si può cercare di risolvere con buone possibilità di successo se si dispone di un contesto della frase e si riesce a padroneggiare il significato complessivo.

J43 r.12 Per elaborare linguaggi naturali è necessario operare non solo sulle frasi scritte (componenti lessicali), ma anche sulle loro strutture sintattiche (ad esempio alberi sintattici alla Chomsky) e sulle corrispondenti componenti semantiche, ossia sui significati.

J43 r.13 Problemi relativi alla traduzione automatica

J43 r.14 La difficoltà di interpretare il parlato spesso viene aggravata dal fatto che molti parlanti non seguono regole chiare, ma usano espressioni poco ocontrollate e qualche volta emettono frasi approssimative o incomplete in quanto accompagnate da messaggi del linguaggio del corpo; il loro testo risulta di difficile interpretazione se non si padroneggia l'intero scenario entro il quale sono state pronunciate. Talvolta sono pronunciate frasi che non seguono un flusso logico per vari motivi, ad esempio psicologici, che un osservatore esterno può giudicare pronunciate a vanvera, prive di un significato rilevabile.
La interpretazione di tutte queste frasi risulta particolarmente difficoltosa per gli algoritmi.

J43 r.15 Il termine "assistente vocale" denota un prodotto software o un dispositivo digitale in grado di interpretare il linguaggio naturale e che, se appositamente "addestrato", può dialogare con interlocutori umani allo scopo di acquisire o fornire informazioni, di portare avanti dei dialoghi su temi convenzionali o dei dibattiti oppure di gestire la effettuazione di vari tipi di operazioni nelle quali tradizionalmente sono appannaggio di operatori umani.

Gli utenti di un assistente vocale possono porgli delle domande e fornirgli istruzioni vocali per scopi quali controllare dispositivi domotici, richiedere la riproduzione di contenuti multimediali, gestire attività quotidiane che riguardano l'e-mail, le agende, i calendari o altre operazioni routinarie.

Gli assistenti vocali spesso sanno rispondere servendosi di sintetizzatori di voci.

Il riconoscimento vocale ha trovato i primi impieghi a partire dagli anni 1990 nei call centers da parte di operatori che utilizzano i suoi strumenti per gestire le chiamate tramite interazioni vocali con il chiamante.
Intorno al 2000, in relazione alla migliorate prospettive tecniche e a nuove potenzialità commerciali, accanto ad alcune aziende storiche del riconoscimento vocale (Nuance), sono entrati in scena i giganti dell'alta tecnologia come Amazon, Apple, Google e Microsoft.

Tra le più famose implementazioni del riconoscimento vocale degli inizi del secolo XXI si ricordano quelle disponibili in Windows XP dal 2001 e in alcuni dei primi cellulari che negli anni 2000 hanno iniziato a permettere la dettatura vocale degli SMS e il riconoscimento dei nomi delle rubriche, consentendo agli utenti di avviare agevolmente le chiamate telefoniche con comandi vocali.

J43 r.16 Una svolta importante è arrivata con il machine learning e l'intelligenza artificiale avanzata che hanno permesso di migliorare drasticamente l'efficienza del riconoscimento vocale rendendo possibile la nascita dei assistenti vocali sui quali si poteva contare con una buona regolrità.

La prima azienda a dare un forte impulso in tal senso è stata Apple, che nel 2011 ha lanciato l'assistente SIRI, ancora oggi presente in tutti i prodotti della società, ed esercitando uno stimolo nei confronti di altre aziende costrette a proporre prodotti con analoghe prestazioni.
Oggi Google propone Assistente Google, Amazon propone Alexa, Microsoft Cortana e Samsung Bixby.

Gli assistenti vocali possono essere integrati in diversi tipi di piattaforme. Per attivarli, devono essere pronunciati una parola specifica o una breve frase. Esempi: Hey Siri, OK Google, Alexa e Hey Microsoft.

Il loro utilizzo solleva problemi relativi alla privacy degli utenti: infatti l'attivazione vocale richiede che il dispositivo sia sempre attivamente "in ascolto" delle parole dell'utente e possa interpretare talune delle sue azioni.

J43 r.20 I linguaggi naturali per oltre 5000 anni hanno costituito il principale strumento per la comunicazione interpersonale, per la organizzazione, la registrazione e la diffusione delle conoscenze.
Questo ha indotto a studiarli da molti punti di vista e a controllarle per molteplici scopi.
In particolare sono stati proposti numerosi linguaggi convenzionali per le diverse tecnologie e in particolare linguaggi per la programmazione dei computers.
Non può sorprendere che siano importanti per l'ecosistema AI.

J43 r.21 Disponiamo di modelli di linguaggio probabilistici basati su n-grammi che si dimostrano in grado di catturare una quantità di informazioni riguardanti un linguaggio naturale e forniscono buone prestazioni per compiti quali identificazione del linguaggio da sue frasi, correzione ortografica, analisi del sentiment, classificazioni secondo generi, riconoscimento di entità identificate da nomi.

Dato che le espressioni adottate in ogni linguaggio naturale posseggono numerose varianti lessicali e sintattiche, oltre a tante sfumature di significato, per trattare i modelli probabilistici dei linguaggi si rende necessario applicare preelaborazioni e regolarizzazioni dei dati per ridurre il "rumore" che accompagna spesso le registrazioni delle loro frasi.

Nel definire un sistema statistico per un linguaggio è consigliabile adottare un modello in grado di far buon uso dei dati, anche se può apparire troppo semplificato e rischia di perdere molti elmenti di dettaglio portatori di effetti distintivi che possono essere rilevanti.

Le rappresentazioni con word embedding possono fornire informazioni assai utili delle parole e dei loro collegamenti per similarità.

È utile catturare la struttura gerarchica di un linguaggio e per questo servono le grammatiche a struttura sintagmatica e in particolare le grammatiche acontestuali (v. a. C14a).
Sono ampiamente usati il formalismo della grammatica acontestuale probabilistica (PCFG) e quello della grammatica delle dipendenze.

J43 r.22 Le frasi di un linguaggio acontestuale possono essere analizzate in un tempo $O(n^3)$ da un parser di chart come l'algoritmo CYK che richiede regole grammaticali nella forma normale di Chomsky

[C14c12]. Con qualche perdita di accuratezza le frasi di linguaggi naturali possono essere analizzati in tempi $O(n)$ servendosi di una ricerca beam o di un parser del tipo shift reduce.

Per l'apprendimento di una grammatica PCFG e dei relativi parametri può risultare utile un treebank.

È notevolmente utile aumentare una grammatica arricchendola di informazioni e richieste utili per gestire aspetti come la concordanza soggetto-verbo, le funzioni dei pronomi e per rappresentare informazioni su singole parole e non solo su intere categorie di parole.

Una grammatica aumentata può anche supportare la interpretazione semantica.
È possibile apprendere una grammatica con supporto semantico a partire da un corpus di domande dotate di spiegazioni, di caratteristiche logiche o di risposte.

Ogni linguaggio naturale oggi può essere trattato con molti strumenti formali, ma rimane complesso e difficile da rappresentare con i formalismi. Tra questi vi sono gli strumenti delle attività dell'area deep learning.

J43 r.23 Le rappresentazioni continue delle parole mediante word embedding risultano sensibilmente più robuste delle rappresentazioni discrete che trattano le parole come atomi. Esse infatti possono essere preaddestrate con data sets di testi non etichettati.

Con reti neurali ricorrenti si possono modellare efficientemente il contesto locale e quello a lunga distanza, mantenendo informazioni rilevanti nei loro vettori di stato nascosto.

Sono stati messi a punto modelli sequenza-sequenza utilizzabili per la traduzione automatica e per la generazione di frasi e testi.

Sono studiati i cosiddetti modelli transformer che usano la auto-attenzione e riescono a modellare il contesto locale e il contesto di lunga distanza. Essi possono trarre vantaggio anche da operazioni di moltiplicazione matriciale, oggi eseguibile da hardware specializzato con grandissima efficienza.

L'apprendimento per trasferimento che include rappresentazioni word embedding contestuali preaddestrate consente di sviluppare modelli a partire da corpora molto estesi, anche privi di etichette e di applicarli ad ampi ventagli di compiti.

I modelli preaddestrati per predire parole mancanti in testi incompleti riescono ad eseguire altri compiti come risposte a domande e implicazione testuale richiedendo solo una fase di precisazione del dominio di destinazione.

J43 s. Big data

J43 s.01 Il termine big data viene usato per comprendere le molte attività che riguardano le utilizzazioni delle grandissime quantità di dati e che le recenti tecnologie consentono di raccogliere, gestire ed elaborare. Queste utilizzazioni sono considerate del massimo interesse grazie alla loro elevata potenzialità statistica nei confronti del miglioramento delle conoscenze delle realtà e dei comportamenti delle popolazioni.

Le masse di dati ora disponibili risultano troppo estese e complesse per essere trattate con il software per la gestione delle basi dati che si era consolidato negli anni dal 1974 al 2000 con le garanzie fornite dal modello relazionale di Codd e dalla strumentazione ICT in risposta alle esigenze prevalentemente amministrative di quel periodo.

Si dice anche che si giunge al livello dei big data quando per manipolare i dati per le applicazion si rendono necessari strumenti di parallel computing. Per queste attività si usano anche molte migliaia di servers e recentemente si sono adottati sistemi di micro elettronica e anche di nanoelettronica costruiti su wafers.

Le attività che afferiscono all'area big data sono molteplici: comprendono la raccolta dei dati, il loro immagazzinamento, le ricerche all'interno delle loro collezioni, i monitoraggi statistici, la loro integrazione, le operazioni che consentono modalità versatili di condivisione, la loro distribuzione in genere preceduta da selezione e spesso seguita dalla sua storicizzazione, la loro visualizzazione, gli aggiornamenti del posseduto e del distribuito con cadenze diverse, le ristrutturazioni, la gestione dei diritti di accesso e della privacy e il controllo delle fonti.

J43 s.02 All'inizio ai big data si sono associate le esigenze dette delle 3V: volumi, velocità e varietà; a queste si sono in seguito raggiunte le 5V tenendo conto anche delle esigenze della veracità e del valore. Presentiamo in breve questi requisiti.

Grandi volumi di memoria, al di sopra di 10 TeraByte, ossia di 10^{10} bytes; oggi si è giunti a prospettare la gestione di quantità di memorie intorno allo ZettaByte, cioè ai 10^{21} bytes.

Grandi velocità di elaborazione; se questa mancasse si avrebbero tempi di attesa inaccettabili per molte attività di ricerca, di analisi e di ristrutturazione.

Grande varietà delle strutture dei dati e dei formati dei files che possano essere monitorati, stante la tendenza a controllare tutti i dati raccoglibili.

Elevata veridicità delle fonti dei dati costituenti le collezioni trattate.
Occorre osservare che senza adeguati investimenti volti ad assicurare la veridicità il volume e la varietà dei dati utilizzati conducono a forti rischi di errori nelle applicazioni.

Elevato valore per i possibili utenti dei big data garantito dalla alta capacità di utilizzo dei dati sotto controllo.

Taluni fanno riferimento alla sigla 5VC, estensione della 5V con una C che riguarda il requisito della capacitdi affrontare la complessità dei dati, che in genere provengono da fonti disparate, seguono formati diversi e devono essere armonizzati prima di poter essere analizzati e rielaborati da sistemi di programmi che sarebbe opportuno fossero tendenzialmente omogenei per essere meglio interconnessi.

Altri aggiungono l'esigenza di saper affrontare la variabilità, per tener conto dei continui cambiamenti del mondo odierno che inducono variabilità nei modi di generare nuovi dati e delle esigenze per le quali i dati possono essere consultati e utilizzati.

I big data possono raccogliere dati strutturati, nonstrutturati e mescolanze degli uni e degli altri in quanto provenienti da molteplici fonti.
Per affrontare la variabilità si deve tener conto dei possibili cambiamenti dei formati, delle strutture e delle organizzazioni delle fonti.

J43 s.03 Per le metodologie applicative dei big data occorre innanzi tutto segnalare la predictive analytics, la user behavior analytics e altri metodi di analisi avanzati studiati proprio per estrarre conoscenze dalle grandi masse di dati.

Dai big data si cercano di estrarre nuove correlazioni riguardanti tendenze per le attività dei pubblicitari, prevenzione delle malattie, lotta alla criminalità.
Si pongono impegnativi problemi come quelli che incontrano le ricerche in Internet, quelli che riguardano le notizie che servono alle scelte finanziarie, quelli delle analisi per la cura della salute, quelli che vengono dai sistemi informativi geografici (GIS), dalla informatica urbana e dalla informatica per i commerci.

Nell'ambito della cosiddetta e-science si manifestano esigenze specifiche riguardanti tra l'altro: meteorologia, genomica, connectomica, si rendono necessarie complesse simulazioni per la fisica, per la chimica, l per la biologia e per i problemi ambientali.
Lo studioso Jim Gray, sostiene che la e-science si possa chiamare anche "data-intensive science" e sia il quarto paradigma della scienza che completa la sequenza:

empirica, teoretica, computazionale e (ora) guidata dai dati.

Egli e afferma che "per la scienza tutto sta cambiando per opera della tecnologia dell'informazione" e del diluvio dei dati.

J43 s.04 Le dimensioni e il numero degli insiemi di dati disponibili sono cresciuti rapidamente da quando i dati vengono raccolti da insiemi di strumenti sempre più diffusi come: dispositivi mobili, apparecchiature di telerilevamento da velivoli (droni compresi), dispositivi di rilevamento dell'Internet delle cose tendenzialmente economici, registrazioni dei testi sorgente del software e i logs delle loro esecuzioni, telecamere, microfoni, lettori di identificazione a radiofrequenza (RFID) e reti di sensori wireless.

La capacità tecnologica pro-capite di immagazzinare informazioni dagli anni 1980 è raddoppiata ogni 40 mesi circa; Nel 2012, ogni giorno sono stati generati dati corrispondenti a 2,5 exabytes, ossia a $2,5 \times 10^{18}$ bytes.

In base alle previsioni di un rapporto IDC, si prevede che il volume globale dei dati tra il 2013 e il 2020 abbia avuta una crescita esponenziale da 4,4 zettabyte a 44 zettabyte.
Entro il 2025, IDC prevede che ci saranno 163 zettabyte di dati e che la spesa globale per le soluzioni di big data e business analytics (BDA) si stima che nel 2021 abbia raggiunti i 215,7 miliardi di dollari.

Secondo un rapporto di Statista, si prevede che il mercato globale dei big data crescerà a 103 miliardi di dollari entro il 2027.
Nel 2011 McKinsey & Company ha dichiarato che se la sanità statunitense utilizzasse i big data in modo creativo ed efficace per migliorare la sua efficienza e la sua qualità, il settore potrebbe generare un valore di oltre 300 miliardi di dollari all'anno.

Nelle economie sviluppate dell'Europa, gli amministratori pubblici potrebbero risparmiare più di 100 miliardi di euro in miglioramenti dell'efficienza operativa grazie al solo accurato utilizzo dei big data.
Inoltre gli utenti dei servizi abilitati dai dati di localizzazione personale potrebbero ottenere 600 miliardi di dollari di surplus per i consumatori.

Dopo queste considerazioni le grandi imprese si pongono il problema di come si stabilisca chi debba essere il responsabile delle iniziative sui big data che riguardano l'intera organizzazione della loro gestione

J43 s.05 Per architetture specifiche v. we we.
Per le tecnologie specifiche si ricorre a tecniche di analisi dei dati, come A/B testing, machine learning, ed elaborazione dei linguaggi naturali.
Le tecnologie dei big data interagiscono intensamente con la business intelligence, il cloud computing e la gestione delle basi dati.
Per le visualizzazioni vengono adottate carte, grafi e altri dispositivi consolidati per queste attività.
Inoltre si ricorre sistematicamente all'Online analytical processing (OLAP), al massively parallel-processing (MPP) effettuabile dai supercomputers e dalle computing farms , al data mining, ai sistemi di files distribuiti, al burst buffer e al Memcached.

J43 s.06 La crescita dei big data e dei sistemi per la loro gestione pongono importanti esigenze ai settori dell'educazione e dei media.

La strumentazione IoT viene adottata sempre più estesamente, anche come mezzo per raccogliere dati sensoriali, dati che vengono utilizzati sempre più ampiamente negli ambiti medico, manifatturiero e dei trasporti.

J43 s.07 Sono numerose le applicazioni nelle quali le nuove tecnologie dei big data si sono dimostrate determinanti.
Quando il Large Hadron Collider del CERN è stato impegnato nella individuazione del bosone di Higgs, sono state richieste misurazioni che partono da dati raccolti primariamente al ritmo di 600 milioni di collisioni per secondo; da queste vengono estratte 1000 collisioni al secondo potenzialmente interessanti e dopo questa riduzione rimangono da analizzare flussi di 200 Pbyte all'anno.

Lo Square Kilometre Array è un radiotelescopio costituito da migliaia di antenne. Si prevede che dal 2024 saranno prodotti 14 exabyte al giorno che saranno ridotti a 1 Pbyte al giorno da conservare per essere sottoposti ad esami successivi.

Il telescopio Synoptic Survey Telescope produce 140 Tbyte di dati ogni 5 giorni.

J43 s.08 La prima decodifica del genoma umano ha richiesto 10 anni, mentre ora una decodifica viene effettuata in meno di un giorno.

Il Center for Climate Simulation (NCCS) della NASA immagazzina 32 petabytes di osservazioni climatiche.

L'azienda DNAStack di Google compila e organizza campioni di DNA ricavati da dati genetici provenienti dalle varie parti del pianeta; questi campioni servono ad identificare malattie e altre anomalie di interesse medico sanitario e per questo sono da sottoporre ad elaborazioni molto inpegnative.

La base dati 23andme's DNA contiene le informazioni genetiche di un milione di persone delle varie parti del mondo.

L'archivio Johns Hopkins Turbulence Databases (JHTDB) contiene più di 350 terabyte di dati che caratterizzano domoni spaziotemporali che servono a tenere sotto ontrollo la grande varietà dei fenomeni collegati alla turbolenza.

J43 s.09 I big data sono impiegati ampiamente per sostenere alcune attività sportive.

Vi sono attività che si servono di grandi quantità di informazioni per definire i protocolli degli allenamenti di diverse tipologie di atleti.

Le case automobilistiche impegnate nelle gare di Formula Uno raccolgono ed analizzano grandi quantità di dati in gran parte ottenute da sensori montati sulle vetture nel corso delle prove e delle gare. Questi dati vengono utilizzati quasi in continuo per la messa a punto dei motori, degli impianti interni delle vetture e dei pneumatici.

Nelle varie fasi della pandemia da Covid-19 sono state raccolte grandissime quantità di dati in numerose parti del mondo per cercare di tenere sotto controllo le diffusioni delle varianti, gli effetti degli interventi medici, soprattutto di quelli prodotti dalla somministrazione dei vaccini, e i confronti tra le diverse iniziative sanitarie.

J43 s.10 Sono enormi le quantità di dati raccolte dalle iniziative sul Web, soprattutto quelle dei cosiddetti giganti del Web.
eBay ....
Amazon ....
Facebook ....
Google ....

J43 s.11 Critiche ...

J43 t. Visione artificiale

J43 t.01 La percezione ha lo scopo di ricavare dallo spazio che circonda il percepente le informazioni che gli sono necessarie per scelte operative quali spostarsi, navigare, riconoscere oggetti, stabilire relazioni fra oggetti, manipolare cose e dispositivi.

Essa quindi richiede che il percepente effettui operazioni che in molte circostanze si rivelano impegnative.

La geometria e l'ottica, che sono alla base della comprensione della formazione delle immagini acquisite con la vista, sono state ampiamente studiate e i loro risultati sono di aiuto allo sviluppo degli strumenti della visione artificiale.

La grafica aveva già affrontato il problema consistente nel disporre di una descrizione sufficiente di una scena 3D e di produrre una sua immagine 2D relativa a un qualsiasi punto di vista: questo compito lo sappiamo svolgere abbastanza facilmente anche per il fatto che le relative indagini hanno natura strettamente deduttiva.

È più impegnativo il problema inverso, di natura induttiva, quello della visione artificiale consistente nell'esaminare una immagine 2D (o più d'una) e di ricavarne una descrizione di una scena 3D.

J43 t.02 Le rappresentazioni utili delle immagini è opportuno che consentano di individuare bordi, tessiture, flusso ottico e regioni spaziali; questi elementi costituiscono importanti indizi per individuare conoscenze quali confini di oggetti e e corrispondenze tra immagini diverse di una stessa scena.

Le reti neurali convoluzionali sono in grado di ricavare dalle immagini caratteristiche che consentono di individuare accurati elementi classificatori delle immagini stesse.
Dal punto di vista metodologico queste caratteristiche in genere si possono considerare patterns di patterns di patterns.

Purtroppo non è facile prevedere in quali applicazioni questi classificatori funzioneranno, cioè soddisferanno le esigenze, in quanto le immagini di addestramento del sistema classificatore potrebbero avere caratteristiche diverse da quelle della applicazione per qualche aspetto importante ma difficile da individuare.
È l'esperienza di queste attività che può aiutare a capire quando un classificatore disponibile può risultare accurato in uno specifico contesto applicativo.

J43 t.03 La visione artificiale è un processo complesso di grande importanza per la robotica e per la costruzione di tutti i dispositivi che consentono di avvalersi delle informazioni fornite da immagini tratte dall'ambiente nel quale sono posizionati.

Per questi studi vengono adottati modelli dei tipi di oggetti definiti da procedimenti consolidati presi come riferimenti, in particolare da sistemi CAD e da modelli di rendering che hanno come fine la precisazione dei processi geometrici, fisici e statistici che producono i segnali che l'agente percepisce.
Le due problematiche fondamentali della visione artificiale sono la ricostruzione, con la quale l'agente definisce un modello della porzione di ambiente presa in visione a partire da una o più immagini, e il riconoscimento con il quale l'agente effettua la distinzione tra gli oggetti da cui riceve informazioni visive o d'altro genere (tattili, sonore, ...).

J43 t.04 Vengono percepite immagini senza utilizzo di lenti (stenoscopio = pinhole camera) o con la mediazione di lenti, come accade servendosi di sistemi ottici e come fanno gli stessi occhi dei viventi.

Può essere necessario tenere in debito conto di svariati effetti: effetti di prospettiva, riflessioni, luce ambiente, illuminazione funzionale, ombre e colori.
Per questi ultimi si fa riferimento al principio di tricromia (dovuto a Thomas Young) che, in accordo con il fatto che l'occhio umano dispone di tre tipi di ricettori, considera che ogni colore sia scomponibile in tre componenti: rosso, verde e blu, ossia red, green, blu.
Si tratta quindi della scomposizione RGB dei segnali raccolti dai sensori, ciascuno dei quali chiamiamo **pixel**, picture element, e in genere valutiamo con 3 bytes ciascuno con una intensità espressa con una coppia di cifre esadecimali da 00 a FF, ossia con interi decimali da 0 a 255.

La luce che costituisce una immagine raccolta da un agente è la luce che viene riflessa dagli oggetti che contribuiscono all'immagine; tipicamente una immagine oggi viene costituita da 12 milioni di pixels e quindi richiede una quantità considerevole di dati.

J43 t.05 I classificatori di immagini possono essere adoperati per gli algoritmi rilevatori di oggetti. Un classificatore viene usato per classificare alcune parti (boxes) di una immagine attribuendole un punteggio della qualità chiamata "onjectness". Un altro decide se un box contiene uno specifico oggetto o un oggetto di uno specifico campionario.

I metodi di rilevamento attuali possono commettere errori, ma possono rivelarsi utili in varie applicazioni.

Non sono molte le situazioni nelle quali da una sola vista si riesce a ricavare la geometria 3D. Disponendo di più viste invece è spesso possibile ricavarla.

J43 t.06 Le applicazioni della visione artificiale complessivamente sono molto numerose e varie.

Da una decina di anni si parla anche di "machine vision", attività afferente all'ambito industriale che riguarda tecnologie e metodi per estrarre informazioni da immagini in modo automatico.
Questo genere di attività diffusa è stato reso possibile dai successi della visione artificiale con metodi di deep learning e viene praticata per la ispezione e l'ordinamento automatici basati su immagini e per la guida dei robots.
Le attività di ispezione e ordinamento richiede acquisizione dell'immagine, elaborazione digitale dell'immagine comprendente operazioni di machine learning mediante reti neurali e deep learning, pattern recognition, OCR, misurazioni e decisioni finali che possono concretarsi in una scelta accetto/rifiuto o in una decisione di classificazione.

J43 t.07 Per analizzare i dati esprimenti un'immagine si devono ottenere sue rappresentazioni in qualche misura semplificate. Evidentemente le rappresentazioni più sono semplificate, più perdono dettagli e quindi dovrebbero continuare a contenere elementi di elevata influenza.

Attualmente si prendono in maggiore considerazione quattro proprietà giudicate di importanza generale: bordi, tessitura (texture), flusso ottico o optical flow e segmentazione in regioni.

I bordi sono linee dell'area immagine che delimitano significativi cambiamenti di luminosità.

Spesso si ottiene applicando uno smoothing, un allisciamento dei dati al fine di ridurre il rumore con un filtro gaussiano sottoposto a una convoluzione.

Il termine texture indica un pattern, cioè una configurazione visiva riscontrabile in una regione dell'area immagine; esempi ciotoli di una spiaggia, fili d'erba in un prato, persone in una folla, macchie sul manto di un felino, onde del mare.
Per questi patterns si usa anche il termine **texel**.

J43 t.08 Per flusso ottico si intende il movimento apparente degli oggetti che compaiono in una sequenza video ricavata da una macchina in movimento. Si rappresenta matematicamente con il campo vettoriale delle velocità dei dettagli puntiformi rilevati e si riconoscono le somiglianze tra dettagli in movimento con parametri come la somma dei quadrati delle differenze, ossia sum of squared differences, in sigla SSD.

La segmentazione dell'immagine in regioni si basa sulla distinzione tra piccole o grandi variazioni delle proprietà visuali degli oggetti o dei patterns individuati.
Si effettua o dando importanza ai confini tra le zone con proprietà diverse, o procedendo a raggruppare i pixels con caratteristiche simili e i piccoli aggregati di pixels con aspetti simili e successivamente riuscendo a individuare raggruppamenti dei suddetti piccoli aggregati ai quali si applica il termine di superpixels. Questi consentono una prima segmentazione dell'immagine.

J43 t.09 Un'attività molto importante è la classificazione delle immagini.
Se in una immagine si rilevano degli oggetti ci si avvale di tassonomie di oggetti attribuiti a loro categorie.
Il riconoscimento dell'aspetto degli oggetti (colore e texture) può incontrare varie difficoltà imputabili a:
diversa illuminazione che porta a diversa luminosità e diversi colori;
diverso scorcio, ossia diverso angolo di illuminazione o di riflessione;
diverso punto di vista;
possibile occlusione o autoocclusione, ossia nascondimento di parti di un oggetto o di un pattern;
possibile deformazione dovuta a eventi non sufficientemente dominati.

J43 u. Robotica

J43 u.01 Una definizione un poco astratta afferma che la robotica si occupa di agenti dotati di un corpo fisico che sono in grado di modificare la parte del mondo fisico nel quale vengono collocati, il loro cosiddetto workspace.

In termini più operativi la robotica si definisce come la disciplina che si colloca tra l'informatica e l'ingegneria meccanica e riguarda la progettazione, la costruzione, la messa in opera e l'utilizzo dei robots (o robota), macchine in grado di effettuare una certa gamma di azioni eseguibili dalle persone e in grado di sostituire vantaggiosamente gli operatori umani in varie situazioni prevedibili.

Attualmente i robots vengono usati largamente per sostituire l'uomo in attività classificabili come segue.
Lavori che richiedono grandi sforzi o grande velocità di intervento.
Manovre molto difficili da controllare come quelle che comportano l'esecuzione di micromovimenti (robots chirurghi).
Manovre che chiedono di effettuare scelte basate su molte caratteristiche in tempi brevissimi (come accade in molte emergenze).
Attività pericolose (sminamento, manipolazione di materiali radioattivi o velenosi, azioni rischiose).
Attività da svolgere in ambienti ostili, ambienti nei quali l'uomo non opera in sicurezza e al limite ambienti nei quali non riuscirebbe a sopravvivere. In particolare questo accade in veicoli spaziali impegnati in missioni di lunghissima durata, per i quali non è previsto il ritorno o di lunghissima durata o in condizioni fisiche estreme; questi veicoli devono essere robotizzati, cioè alla loro conduzione devono essere incaricati dei robots adeguatamente resistenti.

J43 u.03 La tipologia dei robots è ormai decisamente ampia e continua a crescere.

Una importante distinzione dei robots riguarda la loro consistenza fisica.

I manipolatori, o bracci robotici, sono meccanismi che dispongono di sensori con i quali esaminano il loro workspace e possiedono attuatori in grado di agire su questa zona di spazio.

I robots mobili utilizzano gambe meccaniche, ruote o eliche per muoversi. Oggi sono disponibili droni quadricotteri e droni con più di 4 eliche, veicoli senza pilota, automobili con un robot che aiuta il guidatore e si alterna alla guida (in particolare in autostrada), veicoli subacquei autonomi e rovers per le esplorazioni di corpi celesti a noi vicini: Luna, Marte e altri satelliti del sistema solare.

Altri robots riguardano protesi, esoscheletri, robots alati, sciami e interi locali abitabili.

J43 u.04 In genere i robots sono dotati di sensori passivi poco energivori.
Vi sono comunque robots dotati di sensori attivi che possiamo dire dialogano con il loro workspace.

Tra i sensori dei robots si incontrano telemetri, fotocamere che forniscono contributi alla stereoscopia, lidar all'interno di automobili robotizzate, videocamere capaci di valutare tempi di volo, radar, sensori tattili, sensori di posizione come GPS e GPS differenziale, sensori propriocettivi, odometri per misurare distanze percorse a partire da movimenti delle ruote, sensori inerziali, sensori di forze, sensori di torsioni.

I robots si possono servire di attuatori elettrici, idraulici e pneumatici; un robot può disporre di giunti rotanti, di giunti prismatici con un solo asse di movimento e anche di giunti a più assi.

Per afferrare un robot si serve di pinze a due o tre ganasce oppure di mani umanoidi.

J43 u.05 Molte le possibili caratteristiche del software.

Sono disponibili programmi che operano deterministicamente e altri che agiscono stocasticamente.
Vi sono programmi che si servono di funzioni di ricompensa.
Le applicazioni nelle quali agiscono più robots si servono di il software che fa riferimento alla teoria dei giochi.

Per molte scelte operative si distinguono un livello di pianificazione dei compiti, un livello di pianificazione del movimento e un livello di controllo.

Talora un robot deve ricevere un apprendimento delle preferenze utili a stimare un obiettivo finale; altre volte serve la predizione dei comportamenti degli umani che dovrà servire.

Quando si possono presentare diverse aree di comportamenti si pone il problema generale dell'integrazione delle preferenze per le diverse aree.

La percezione dei robots pone problemi di stima del proprio stato che a sua volta concerne informazioni che consentono di prendere decisioni buone, aggiornabilità delle stesse e buona corrispondenza tra variabili interne e variabili della zona del mondo fisico che lo attornia.

Ad alcuni robots servono azioni di mappatura degli oggetti fisici da coinvolgere e modelli sensoriali che tengono conto di riferimenti spaziali o landmarks.
Si pongono quindi problemi di SLAM, simultaneous localization and mapping per i quali possono servire filtri di Kalman estesi.

J43 u.06 I robots possono anche percepire temperature, pressioni, suoni e odori.
Per la percezione robotica possono porsi problemi di apprendimento supervisionato o meno.
Molti robots adottano tecniche di percezione adattiva.
Servono pianificazione del movimento e controllo dell'inseguimento della traiettoria.

Possono servire grafi di visibilità che nel semplice caso bidimensionale si avvalgono dei grafi di Voronoi.
In alternativa si adotta la suddivisione del workspace in celle e si opera con una funzione binaria che fa da rilevatore di collisioni.

J43 u.07 Passiamo in una rapida rassegna i domini applicativi dei robots.

Assistenza a domicilio ; in questo ambito si studiano interfacce cervello-macchina per tetraplegici.
Sempre nell'ambito della sanità va ricordato l'aiuto ai chirurghi.

Ai robots si affidano sempre più servizi di uffici, alberghi e ospedali; molti robots sono incaricati di servizi di telepresenza.
Automobili autonome (driver assist per la guida in autostrada).
Intrattenimento (animatronics e robot della Disney)
Esplorazioni e ambienti pericolosi
Industria oggi la maggior parte
I problemi della robotica riguardano temi come la stocasticità V. RNii332

J43 u.08 Rassegna degli elementi caratterizzanti i robots
Struttura meccanica [A]
Componenti elettronici [A]
Software incorporato [A]
Componenti generatori di energia: batterie, motori a combustione, gas compressi, energia solare, flywhhel energy storage, (energia nucleare)

Attuatori di movimenti: soprattutto motori elettrici, attuatori lineari pneumatici e idraulici, attuatori elastici, muscoli ad aria, muscoli da elettrocontrazione, polimeri elettroattivi (EAP), piezomotori, nanotubi elastici.
Sensori per la visione, per il tatto e per l'udito [A] Sensori di posizione, di velocità

J43 u.09 Presentiamo una rapida panoramica dei vasti campi di applicazione dei robots.
Robots militari
Robots industriali
Robots collaborativi (Cobots)
Mani robotiche
Esoscheletri robotici
Robots agricoltori
Robots per la cucina
Robots lottatori
Robots decontaminanti
Robots domestici

J43 u.10 Esaminiamo con qualche dettaglio in più le applicazioni dei robots nel mondo dell'assistenza medica e sociale.

Medicina di precisione.

iCub, robot bambino Open Source prodotto da IIT ccon la prospettiva di utilizzarlo da parte di neuropsichiatri per sostenere pazienti affetti da autismo.

Progetto Blissino
J43 u.11 Aspetti etici e psicologici della robotica

J43 v. Machine learning

J43 v.01 Il termine machine learning, tradotto con apprendimento automatico, si riferisce a un complesso di procedimenti sviluppati in gran parte a partire dagli ultimi anni del secolo scorso che in comune hanno lo scopo di migliorare le prestazioni dei vari algoritmi finalizzati al riconoscimento di patterns di un determinato tipo attraverso qualche genere di esame statistico dei risultati dell'applicazione a casi simili degli algoritmi stessi.

Dal punto di vista dei programmatori l'apprendimento automatico è un potenziamento della programmazione tradizionale (che invece si può caratterizzare con l'atteggiamento deterministico) attraverso operazioni che si possono decrivere come finalizzate all'apprendimento di caratteristiche dei dati che non sono derivabili da proprietà dei dati stessi conosciute fin dall'inizio del loro trattamento.

Nel machine learning confluiscono molti metodi associabili all'area dell'intelligenza artificiale che sono studiati anche autonomamente in relazione a esigenze relativamente circoscritte.
Per caratterizzare questi metodi sono indicativi i termini statistica computazionale, riconoscimento di patterns, reti neurali artificiali, filtraggio adattivo, elaborazione delle immagini, data mining e algoritmi adattivi.

I procedimemti per l'apprendimento automatico oggi trovano appplicazione in una grande varietà di problematiche: medicina, filtraggio delle e-mail e di altri generi di materiali che circolano sul Web, individuazione di intrusi che cercano di violare archivi, OCR, riconoscimento vocale, visione artificiale, esami delle informazioni finanziarie, analisi predittiva per iniziative commerciali, ...

J43 v.02 Diamo una definizione operativa dell'apprendimento che hanno la possibilità di essereposti in atto da opportuni programmi per il computer.

Consideriamo un programma $\mathcal{P}$ costruito per eseguire un insieme di compiti (tasks) T attraverso la elaborazione di dati appartenenti a un insieme D; più precisamente diciamo che trasformando il dato, prevedibilmente composito, $D \in \mathsf{D}$ esegue il compito $T = \tau(D)$ e che si ha la trasformazione complessiva $\mathsf{T} = \{D \in \mathsf{D} :| \tau(D)\}$.

Diciamo che il programma $\mathcal{P}$ è in grado di effettuare apprendimento se ha la capacità di osservare i risultati ottenuti in seguito alla esperienza E consistente nell'attribuire valutazioni $v(D)$ alle proprie prestazioni $P(D)$ per le elaborazioni dei dati $D \in \mathsf{D}_E \subset \mathsf{D}$ in relazione al maggiore o minore soddisfacimento dei compiti $T(D)$ e inoltre se è in grado di modificare i suoi comportamenti in modo da riuscire ad ottenere prestazioni migliori.

Questa definizione esplicitata da Tom M. Mitchell nel suo libro del 2006 "The discipline of Machine Learning" segue il suggerimento di Turing che nel suo articolo "Computing Machinery and Intelligence" del 1952 propone di sostituire la domanda "Le macchine possono pensare?" con la domanda "Le macchine possono fare quello che noi (in quanto esseri pensanti) riusciamo a fare?" Questa è la domanda ha dato spunto al famoso test di Turing.

J43 v.03 Una macchina per realizzare apprendimento deve essere in grado di generalizzare la propria esperienza attraverso operazioni che implementano ragionamenti induttivi e/o abduttivi.

Tale macchina deve adattarsi per eseguire più efficientemente compiti già affrontati e per eseguire compiti nuovi dopo aver fatto esperienze su insiemi di dati che vengono chiamati dati o esempi di addestramento, ossia training examples.

Questi esempi si devono ricavare da una distribuzione probabilistica attribuita secondo un determinato modello all'insieme delle situazioni (compiti) che si intendono affrontare.
Quindi un programma $\mathcal{P}$ deve essere capace di dotarsi di un modello probabilistico riuscendo a prevedere il miglioramento e/o l'ampliamento delle proprie prestazioni.

Degli algoritmi di apprendimento automatico e dell'analisi delle loro prestazioni si occupa quella che chiamiamo teoria computazionale dell'apprendimento.

Va anche segnalato che non vi sono garanzie a priori di consistenza matematica sulla efficacia dei modelli probabilistici e degli algoritmi che ne ricavano apprendimento.

La teoria dell'apprendimento affronta anche la complessità (soprattutto temporale) degli algoritmi, giudicando fattibili solo quelli che richiedono tempi (e spazi di memoria) polinomiali e considerando intrattabili tutti gli altri.

J43 v.04 Gli approcci all'apprendimento automatico vengono distinti in tre categorie:

Apprendimento supervisionato: si forniscono al modello degli esempi costituiti da coppie della forma

$$\langle \text{ possibile input , output desiderato } \rangle$$

e si cerca di ricavare una regola che a ogni input prevedibile associ l'output corretto.

Apprendimento non supervisionato: il modello deve trovare una struttura negli input forniti, privi di qualsiasi qualificazione.

Apprendimento per rinforzo: si hanno sessioni in ciascuna delle quali il modello interagisce con un operatore (che assume il ruolo di insegnante) che è in grado di indirizzarlo verso il raggiungimento di un obiettivo concreto (per esempio guidare un veicolo verso una data meta) e lo avverte quando l'obiettivo fosse raggiunto.

Un'altra strategia consiste nel far apprendere come giocare un gioco attraverso partite effettuate contro un operatore avversario. (computer che addestra se stesso)

J43 v.05 Un'altro genere ipo di modo di apprendere si basa sull'attribuzione degli outputs delle trasformazioni degli inputs in due o più classi e chiede di attribuire ciascun altro input a una delle suddette classi. Questo è il modo seguito per lo sviluppo di filtri anti-spam delle emails in arrivo.

Consideriamo le applicazioni nelle quali si vogliono controllare processi caratterizzati da parametri valutativi continui.

Spesso si adottano procedimenti di regressione, con approccio supervisionato: questo accade, ad esempio, per la valutazione del flusso in un oleodotto attraverso la misura della attenuazione di fasci di raggi gamma che vengono fatti attraversare tale condotto.

In alternativa si adottano procedimenti di clustering con la suddivisione degli input in diversi gruppi; per questi, contrariamente alle procedure della statistica tradizionale, i gruppi non sono predefiniti e quindi si segue un approccio non supervisionato. Questo è il caso dell'esame dei comportamenti dei frequetatori di un sito web.

J43 v.06 La stretta relazione tra machine learning e statistica ha portato a proporre vari punti di vista ibridi.

Facciamo riferimento alla "data science" da intendersi come disciplina che ha tra i suoi scopi quello di raggruppare le attività di analisi e controllo delle grandi quantità di dati.
Vengono qualificate le modalità di modellizzazione statistica dei dati distinguendo i modelli basati sui dati da quelli basati sugli algoritmi (come quelli che consistono nella costruzione di foreste casuali.

Vengono anche portati avanti studi statistici nati negli ambienti del machine learning e per queesti si parla di apprendimento statistico.

Sono individuabili due approcci distinti: reti neurali e apprendimento per rinforzo e utilizzo di una rete progettata per imparare comportamenti specifici.

J43 v.07
Apprendimento automatico e data mining

J43 v.08 Apprendimento automatico e ottimizzazione

J43 v.09 Apprendimento automatico e soft computing

J43 v.10 Apprendimento automatico e logica induttiva

J43 v.11 Apprendimento automatico e alberi di decisione

J43 v.12 Apprendimento automatico e regole di associazione

J43 v.13 Apprendimento automatico e reti neurali

J43 v.14 Apprendimento automatico e programmazione genetica

J43 v.15 Apprendimento automatico e reti bayesiane

J43 v.16 Apprendimento automatico e macchine a vettori di supporto

J43 v.17 Apprendimento automatico profondo e

J43 v.18 Apprendimento automatico profondo e

J43 v.19 L'apprendimento automatico solleva varie problematiche etiche.

Occorre preoccuparsi che i sistemi in grado di decidere non siano addestrati con insiemi di dati faziosi o pregiudizievoli; quasti sistemi se adottano i dati incautamente rischiano di ripresentare faziosità e pregiudizi su larghissima scala.

Con la diffusione di sistemi di machine learning possono venire rafforzati pregiudizi culturali, in particolare il razzismo istituzionale, il classismo e i pregiudizi xenofobi.
Di conseguenza la raccolta responsabile dei dati di training va considerata un aspetto critico dell'apprendimento automatico.

In ragione della onnipresente possibile ambiguità nei linguaggi naturali, le macchine addestrate su gran parte dei corpi linguistici necessariamente rischiano di adottare forme di ambiguità.

J43 w. Apprendimento da esempi, da modelli probabilistici

J43 w.01 Possiamo dire che un agente AI apprende se è in grado di migliorare le sue prestazioni dopo aver effettuato osservazioni sul proprio workspace.
Le attività di apprendimento automatico o machine learning sono essenziali per gli agenti AI, macchine dotate di programmi come tutti gli attuali computers.

È fondamentale che un agente AI sappia apprendere per due motivi generali. I suoi programmatori non sono in grado di prevedere tutte le possibili situazioni nelle quali l'agente dovrà prendere decisioni; queste situazioni l'agente le potrà riconoscere grazie ai sensori e agli strumenti di riconoscimento dei quali è dotato.

Inoltre l'agente quando apprende deve consolidare e aggiornare il modello che ha dello spazio nel quale deve operare, modello che utilizza anche per il riconoscimento delle situazioni.

L'agente che apprende può migliorare diverse sue componenti e i miglioramenti e le tecniche per perseguirli dipendono principalmente dai seguenti fattori.
quale componente intende migliorare;
quali conoscenze l'agente già possiede e quale è il suo modello del proprio worksspace;
di quali dati e di quali feedbacks dispone per il lavoro di apprendere.

J43 w.02 L'apprendimento si effettua in modi diversi in dipendenza della natura dell'agente, del componente che intende migliorare e del feedback che può impiegare.

Se il feedback disponibile proveniente da un insegnante o dall'ambiente in esplorazione fornisce valori corretti per gli esempi di input, il problema prende il nome di apprendimento supervisionato.
In questo caso il lavoro si può rappresentare schematicamente come la precisazione di una funzione $h(x)$.

Se questa funzione fornisce valori continui e ordinati (tipicamente dei pesi) viene detta regressione; se invece fornisce una delle categorie previste per i dati (in genere poche) viene chiamata classificazione.

Può essere conveniente apprendere una funzione che non solo si accordi con i dati attuali, ma che abbia anche buone probabilità di accordarsi con dati prevedibili per il futuri: per questo si devono bilanciare concordanza con i dati e semplicità delle ipotesi.

Possono servire gli alberi di decisione, strutture visualizzabili che possono rappresentare tutte le funzioni booleane. L'euristica del guadagno informativo fornisce un metodo efficiente per trovare un albero di decisione semplice e consistente di fronte ai dati.

J43 w.03 Le prestazioni di un algoritmo di apprendimento possono essere visualizzate mediante una curva di apprendimento che presenta l'accuratezza della predizione su un insieme di tests in funzione della ampiezza dell'insieme di addestramento.

Quando si tratta di scegliere tra più modelli la selezione del modello può determinare buoni valori per i parametri multidimensionali (iperparametri) che qualificano i modelli attraverso la convalida incrociata sui dati di validazione.

Una volta scelti i valori degli iperparametri, si può precisare un modello da considerare il migliore utilizzando tutti i dati di addestramento.

Talora gli errori che si riscontrano pesano molto diversamente. Si può precisare una funzione di perdita che indichi la gravità dei diversi errori, in modo da poter minimizzare la perdita su un insieme di validazione.

Si dispone di una teoria dell'apprendimento computazionale che analizza la complessità del campione e la complessità computazionale dell'apprendimento induttivo. Si può anche calibrare un compromesso tra l'espressività dello spazio delle ipotesi e la facilità dell'apprendimento. induttivo.

J43 w.04 Viene spesso adottato il modello della regressione lineare. I parametri ottimi forniti da tale modello possono essere calcolati esattamente oppure possono essere trovati mediante una ricerca a discesa del gradiente, tecnica che si può applicare ai modelli per i quali non si conoscono soluzioni in forma chiusa.

Un classificatore lineare con soglia rigida chiamato Perceptron, molto discusso intorno al 1970, può essere addestrato mediante una semplice regola di aggiornamento dei pesi dovuti all'adattamento a dati che sono linearmente separabili. In casi più complessi questa regola non porta a convergenza.

La regressione logistica sostituisce la soglia rigida del perceptron con una soglia che si manifesta in modo graduale, definita da una funzione logistica che presenta un andamento sinuoso.
La discesa del gradiente funziona bene anche per dati rumorosi che non sono linearmente separabili.

Vengono adottati modelli non parametrici che usano tutti i dati disponibili per effettuare ogni predizione, anche per cercare di riepilogare i dati con pochi parametri.
Tra gli esempi di modelli di questo genere vi sono quelli chiamati nearest-neighbors e la regressione pesata localmente.

J43 w.05 Le macchine a vettori di supporto trovano separatori lineari con massimo margine per migliorare le prestazioni di generalizzazione del classificatore.
I metodi kernel trasformano implicitamente i dati di input in punti di uno spazio di alta dimensionalità dove si può sperare di trovare un separatore lineare anche se i dati originali non sono separabili.

I metodi ensemble, come bagging e boosting, spesso ottengono prestazioni migliori dei metodi individuali.

Nelle attività di online learning si possono aggregare le opinioni di esperti per arrivare sempre più vicini alla prestazione del miglior esperto anche quando la distribuzioni dei dati cambia di continuo.

Nel complesso occorre tenere presente che la costruzione di un buon modello di apprendimento automatico richiede esperienza e consapevolezza dell'intero processo di sviluppo, dalla gestione dei dati alla selezione e ottimizzazione del modello fino alla sua manutenzione continuativa.

J43 w.06 Vediamo ora l'apprendimento da modelli probabilistici.
Esso si serve sia di medie e varianze che di costruzioni di complessi modelli bayesiani. I loro campi di applicazione toccano informatica, fisica, ingegneria, biologia computazionale, neuroscienze, psicologia e influiscono sugli stessi modelli bayesiani.

I metodi di apprendimento bayesiano formulano l'apprendimento come una forma di inferenza probabilistica che utilizza le informazioni per aggiornare una distribuzione di probabilità a priori sull'ipotesi. Questo approccio costituisce un buon modo per implementare l'atteggiamento ispirato dal rasoio di Ockham, ma diventa rapidamente intrattabile al crescere della complessità dello spazio delle ipotesi.

J43 w.07 L'apprendimento basato sull'ipotesi massima a posteriori, MAP hypothesis, seleziona la singola azione più probabile in base a dati precedentemente osservati. Questi metodo utilizza ancora la distribuzione a priori e spesso si rivela più trattabile dell'apprendimento bayesiano completo.

L'apprendimento basato sulla massima verosimigianza sceglie semplicemente l'ipotesi che massimizza la verosimiglianza dei dati: è equivalente a un MAP con distribuzione a priori uniforme. I casi

semplici come la regressione lineare in reti bayesiane completamente osserabili, soluzioni di massima verosimiglianza, possono essere trovate facilmente in forma chiusa. L'apprendimento bayesiano ingenuo è una tecnica particolarmente efficace che scala bene verso l'alto.

Quando alcune variabili sono nascoste, soluzioni locali di massima verosimiglianza possono essere trovate per mezzo dell'algoritmo chiamato expectation maximization.
Le sue applicazioni includono il clustering supervisionato con miscele di gaussiane e l'apprendimento di reti bayesiane e di modelli di Markov nascosti.

J43 w.08 Apprendere la struttura di una rete bayesiana fa parte dell'attività di selezione di modelli. Normalmente per far questo si deve eseguire una ricerca discreta in uno spazio delle strutture.
Risulta necessario adottare qualche metodo adeguato per gestire il compromesso tra la complessità del modello e l'adattamento ai dati.

I modelli non parametrici rappresentano una distribuzione per mezzo di una collezione di punti; in questo modo di fare il numero dei parametri cresce con l'insieme di addestramento.

I metodi nearest-neighbors esaminano gli esempi più vicini a ciascuno dei punti in questione, mentre i metodi kernel costruiscono una combinazione di tutti gli esempi pesandoli in base alla distanza.

J43 x. Apprendimento profondo

J43 x.01 Per deep learning si intende una famiglia di tecniche per l'apprendimento automatico, molto cresciuta dali anni intorno al 2010.
In queste tecniche le caratteristiche del problema da risolvere sono tradotte in circuiti con molti stati e molti elementi di connessione regolabili sui quali si sviluppano le computazioni che portano alle soluzioni e alle decisioni.

I circuiti computazionali vengono chiamati reti neurali artificiali, anche se sensibilmente diverse dalle reti neurali fisiologiche e da quelle dei primi sviluppi dell'AI come Perceptron.

La qualifica deep si riferisce al fatto che gli accennati circuiti computazionali per essere rappresentativi di molte aspetti dei problemi presentano nodi che in genere sono organizzati in molti strati e presentano connessioni complesse e conseguenti percorsi elaborati sui quali si devono svolgere le computazioni.

Questo fatto fa del deep learning l'approccio più adatto per tematiche impegnative quali riconoscimento visuale di oggetti e scene, traduzione automatica, riconoscimento e sintesi vocale e sintesi di immagini. Inoltre l'apprendimento automatico è una componente imprescindibile dell'approccio, ulteriormente elaborato, chiamato apprendimento con rinforzo.

Con le computazioni del deep learning si sono ottenuti importanti successi, spesso raggiunti dopo primi insuccessi e ripetizioni di nuovi tentativi.

I successi si registrano per problemi che coinvolgono oggetti ai quali si devono assegnare molti parametri, tipicamente le immagini; per questi i calcoli possono riguardare moltissimi parametri, anche molti miliardi, e quindi reti neurali con moltissimi nodi e flussi computazionali che non si sanno descrivere in termini umanamente comprensibili.

Occorre dunque riconoscere che i meccanismi efficienti del deep learning non sono del tutto chiari e si deve tenere presente che quando si implementano si devono adottare molte cautele.

J43 x.02 I metodi per le elaborazioni delle reti neurali come la regressione lineare e la regressione logistica consentono di trattare numerose variabili, ma si servono di cammini computazionali brevi con le variabili di input che influiscono separatamente e consentono di individuare solo funzioni e confini distinguibili linearmente; accade invece che molte situazioni reali sono sensibilmente più complesse.

Per trattare le reti neurali del deep learning si usa spesso l'algoritmo di retropropagazione il quale per minimizzare le funzioni di perdita implementa una discesa del gradiente nello spazio dei parametri.

Le reti convoluzionali risultano particolarmente adatte all'elaborazione di immagini e ad altri compiti in cui i dati si possono ben caratterizzare con una distribuzione a griglia.

Le reti ricorrenti sono invece efficaci per attività con elaborazioni sequenziali tra le quali la modellizzazione di linguaggi naturali e la traduzione automatica.

J43 x.03 L'apprendimento per rinforzo riguarda agenti che devono imparare a comportarsi efficacemente in ambienti sconosciuti servendosi solo della loro stessa esperienza attraverso l'esame dei propri successi e insuccessi, utilizzando solo le loro percezioni ed eventuali meccanismi di ricompense occasionali.

Questo approccio può essere considerato come paradigma per costruire sistemi intelligenti con ampie possibilità applicative.

Il genere dei compiti assegnati all'agente determina il tipo di informazione che esso deve essere capace di apprendere.

Un agente rivolto all'apprendimento per rinforzo che può basarsi su un modello acquisisce o precisa un modello di transizione per l'ambiente della forma $\mathcal{P}(s'|s,a)$ e apprende una funzione di utilità $U(s)$.
Un agente di apprendimento per rinforzo senza modello invece deve apprendere una funzione azione-utilità $Q(s,a)$ o una politica $\pi(s)$.

J43 x.04 Le funzioni di utilità possono essere apprese seguendo tre approcci.

La stima diretta della utilità si serve della ricompensa futura totale osservata per un determinato stato come valutazione diretta per l'apprendimento della sua utilità.

La programmazione dinamica adattiva (ADP) apprende un modello e una funzione ricompensa dalle osservazioni e quindi utilizza le interazioni dei valori o delle politiche per ottenere le utilità da considerare ottime o una politica da dichiarare ottima.
ADP fa uso ottimo dei vincoli locali sulle utilità degli stati imposti dalle misure dalla struttura della vicinanza dell'ambiente.

I metodi basati sulle differenze temporali (TD) regolano le stime della utilità in modo che siano più compatibili con quelli degli stati successori.
Possono essere considerati semplici approssimazioni dell'approccio ADP che non richiedono un modello.
Si è trovato che usare un modello appreso per generare simulazioni di esperienze può accelerare l'apprendimento.

Le funzioni azione-utilità, dette anche funzioni-Q, devono essere apprese con l'approccio ADP o con l'approccio TD.
Con l'approccio TD il Q-learning non richiede alcun modello, né per l'apprendimento, né in fase di selezione delle azioni.
Questo semplifica il lavoro, ma può ridurre la capacità di apprendere in ambienti complessi, in quanto l'agente non può simulare i risultati di possibili svolgimenti di azioni.

J43 x.05 Quando l'agente ha il dovere di scegliere ed effettuare azioni già nel corso dell'apprendimento, deve gestire il compromesso tra il valore stimato delle azioni e la possibilità di apprendere nuove informazioni utili.
Trovare una soluzione pienament soddisfacente per il problema dell'esplorazione è impraticabile, ma si può puntare a individuare semplici euristiche che permettano di ottenere risultati ragionevoli.

Un agente esplorativo deve anche agire con prudenza facendo attenzione di evitare situazioni che lo portino a una morte prematura.

In spazi degli stati molto estesi gli algoritmi di apprendimento per rinforzo devono ricorrere a una rappresentazione approssimata di $U(s)$ o di $Q(s,a)$ per potersi avvalere di manovre che coinvolgono la generalità degli stati.
L'apprendimento per rinforzo profondo, che utilizza reti neurali profonde come approssimazioni di funzioni, ha ottenuto notevoli successi di fronte a problemi impegnativi.

La modellizzazione delle ricompense e l'apprendimento per rinforzo gerarchico sono utili per apprendere comportamenti complessi, in particolare quando le ricompense sono sparse e servono lunghe sequenze di azioni per ottenerle.

I metodi di ricerca delle politiche si devono esercitare su una rappresentazione delle politiche stesse; dopo aver assunta una tale rappresentazione dovranno continuare a cercare di migliorarla sulla base delle prestazioni osservate.

La variazione delle prestazioni nei domini stocastici costituisce un problema serio; nel caso di domini simulati si può cercare di risolvere questo problema ricorrendo alla fissazione anticipata di un grado di casualità.

J43 x.06 L'apprendimento per apprendistato procede attraverso l'osservazione del comportamento di esperti e può risultare una efficace soluzione di ripiego quando è difficile specificare una funzione di ricompensa che risulti soddisfacente per le esigenze dellapplicazione.

L'apprendimento per imitazione formula il problema come apprendimento supervisionato di una politica a partire da coppie stato-azione suggerite da un esperto o da esperienze valutate o auspicate come affidabili.

L'apprendimento per rinforzo inverso si propone di inferire informazioni sulla ricompensa da un comportamento che viene suggerito da un esperto.

L'apprendimento per rinforzo continua ad essere una delle aree di ricerca più attive del machine learning.
Essa evita la necessità di costruire manualmente comportamenti e di procedere ad etichettare i grandi dataset richiesti per l'apprendimento supervisionato; in alternativa esso evita di dover tradurre manualmente in programmi le strategie di controllo.

Le sue applicazioni alla robotica promettono di essere particolarmente importanti: infatti in questo campo si devono gestire ambienti continui, rappresentabili con molte dimensioni e parzialmente osservabili e in essi i comportamenti di successo possono essere costituiti da molti milioni di azioni elementari.

Dunque nell'apprendimento per rinforzo vengono adottati molteplici metodi, in quanto finora non si è trovato un approccio decisamente superior a tutti gli altri.

Il problema di scegliere tra metodi basati su modello e metodi senza modello è, in buona sostanza, quello di scegliere il miglior modo per rappresentare la funzione agente; questa scelta appartiene alle fondamenta delle attività dell'intelligenza artificiale.
Infatti questa disciplina si deve basare su quello che si riesce a conoscere.

Come miglior modo per rappresentare le funzioni agente si adotta una rappresentazione delle caratteristiche dell'ambiente nel quale esso opera che si giudica rivelarsi la più adeguata.
Mentre taluni sostengono che la disponibilità di dati di addestramento sufficienti, l'approccio senza modello aspira ad avere successo in ogni campo applicativo, si può obiettare che per ambienti molto complessi non si possano trovare nel mondo conoscibile dati sufficienti per un addestramento adeguato.
Probabilmente con ambienti via via più complessi la disponibilità di modelli diventa sempre più necessaria.

J43 x.07 È opportuno segnalare che le attività nel settore dell'apprendimento per rinforzo sono state accelerate dalla disponibilità di ambienti di simulazione open source che consentono di stimolare la sperimentazione di agenti di apprendimento e la valutazione delle prestazioni.

l'Arcade Learning Environment (ALE) dell'Università di Alberta nel 2013 ha messo a disposizione un framework di questo genere per operare su 55 giochi popolari della piattaforma Atari. Questa fornisce le schermate a pixels di una partita che l'agente tratta come percezioni; inoltre fornisce i punteggi correnti che sono utilizzabili per misurare utilità e risultati.
ALA è stato utilizzato dal team di DeepMind per implemetare l'apprendimento DQN.

A sua volta DeepMind ha sviluppato varie piattaforme open source di agenti di apprendimento e ha aggiunto un componente nel linguaggio Python per ML alla piattaforma di Blizzard chiamata StarCraft IILearning Environment.

La simulazione AI Habitat di Facebook fornisce un ambiente virtuale fotorealistico per attività robotiche all'interno di edifici e la piattaforma Horizon serve per l'apprendimento per rinforzo in sistemi di produzione industriale.

Synthia è invece un ambiente per la simulazione per il miglioramento della capacità di visione artificiale per le automobili a guida autonoma.

OpenAI Gym fornisce ambienti per agenti di apprendimento per rinforzo compatibili con sistemi come Google Football.