Bollettino
Unione Matematica Italiana

Sezione A – La Matematica nella Società e nella Cultura

Sandra Fontani

Identificazione efficiente

Bollettino dell'Unione Matematica Italiana, Serie 8, Vol. **2-A**—*La Matematica nella Società e nella Cultura* (1999), n.1S (Supplemento Tesi di Dottorato), p. 29–32.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=BUMI_1999_8_2A_1S_29_0>

Bollettino dell'Unione Matematica Italiana, Unione Matematica Italiana, 1999.

La matematica nella Società e nella Cultura
Bollettino U. M. I.
(8) 2-A Suppl. (1999), pag. 29-32

Identificazione efficiente.

SANDRA FONTANI

La presente tesi si occupa di Inferenza Induttiva, una disciplina di carattere logico-informatico teorico che studia «problemi di identificazione». L'argomento centrale verte sull'Inferenza Induttiva Efficiente, con enfasi sugli aspetti probabilistici.

1. – Inferenza Induttiva: introduzione generale.

L'Inferenza Induttiva, nata sulla base di idee poste da E.M. Gold nel 1967 [2], è stata sviluppata da molti autori, tra cui D. Angluin, D. Osherson ([1], [5]). In generale viene studiato il problema di riconoscere, identificare, un oggetto o un fenomeno sconosciuto sulla base di un numero finito di informazioni relative ad esso. Di particolare interesse sono i casi in cui gli oggetti da identificare sono funzioni ricorsive o linguaggi (cioè insiemi ricorsivamente enumerabili) appartenenti a una classe di funzioni ricorsive o linguaggi fissata. Più precisamente il problema è il seguente:

(1) Nell'identificazione di una classe $\mathcal{C}$ di funzioni ricorsive, viene fornito l'elenco $f(0)$, $f(1)$, ..., $f(n)$, ... dei valori di una funzione $f \in \mathcal{C}$. Un agente (umano o meccanico), detto *learner*, deve formulare delle ipotesi su f congetturando, a ogni passo, un indice per una funzione parziale ricorsiva. Il learner ha successo se, al limite, le sue congetture si stabilizzano su un indice per f.

(2) Nel caso dell'identificazione di una classe $\mathcal{C}$ di linguaggi, viene fornito un elenco degli elementi di un linguaggio $L \in \mathcal{C}$ in ordine qualunque. Un learner deve formulare delle ipotesi su L congetturando, a ogni passo, un indice per un insieme r.e.. Il learner ha successo se, al limite, le sue congetture si stabilizzano su un indice per L.

È importante osservare che:

(a) A ogni passo n il learner possiede solo una quantità finita di informazioni: se l'oggetto incognito è una funzione ricorsiva f (un linguaggio L) disporrà dei valori $f(0)$, $f(1)$, ..., $f(n)$ (di $n+1$ elementi di L). Quindi, a ogni passo, il learner non può essere certo che la sua congettura sia giusta, dato che può essere smentita al passo successivo. Per questo l'identificazione può, di solito, essere possibile solo al limite.

(b) Nel caso dell'identificazione di funzioni, il learner riceve implicitamente anche informazione negativa (da un elenco degli elementi del grafico di f si può desumere un elenco degli elementi che non gli appartengono), mentre nel caso di linguaggi il learner ha solo informazione positiva (elementi di L). Esiste però un altro tipo di identificazione di linguaggi (*identificazione su informant*) nel quale il learner riceve informazione negativa (vengono forniti i valori della funzione caratteristica di L).

(c) Di regola, lo spazio delle ipotesi (l'insieme di indici da congetturare per indovinare ogni funzione/linguaggio di una data classe $\mathcal{C}$) è limitato e noto al lear-

ner. Ossia il learner ha «a priori» l'informazione che la funzione (il linguaggio) da identificare appartiene a $\mathcal{C}$. In alcuni paradigmi di identificazione, si richiede anche che il learner congetturi sempre un'ipotesi appartenente ad uno spazio di ipotesi dato all'inizio. Questa richiesta sarà di fondamentale importanza nell'identificazione efficiente.

(*d*) Nel caso dell'identificazione di funzioni, se si trascurano problemi di efficienza, l'ordine con cui vengono dati al learner i valori di una funzione f è irrilevante. Se, ad esempio, il learner ha bisogno di conoscere i primi mille valori di f per poter esprimere una congettura esatta, egli può aspettare finché (magari dopo un numero enorme di passi) li conosce tutti. Nel frattempo avrà ricevuto anche altri valori successivi, ma, se vuole, può ignorarli. Inoltre, sa che prima o poi riceverà tutti i primi mille valori. La situazione è analoga nell'identificazione di linguaggi su informant. Nel caso dell'identificazione di linguaggi solo su dati positivi, l'ordine è invece molto importante. Consideriamo ad esempio due elenchi di ω così definiti:

$$(1)\ 0, 1, 2, 3, \ldots\ldots \quad (2)\ \underbrace{0, \ldots, 0}_{10}, \underbrace{1, \ldots, 1}_{100}, \underbrace{2, \ldots, 2}_{1000}, \underbrace{n, \ldots, n}_{10^{n+1}}, \ldots\ldots$$

È abbastanza probabile che il learner indovini ω con il primo elenco, ma che sbagli con il secondo. Ad esempio, dopo aver visto ripetere 10 per 10^{11} volte, il learner può essere indotto a credere che l'elenco riguardi l'insieme $\{0, 1, \ldots, 10\}$. Secondo gli usuali paradigmi di identificazione, un learner ha successo se identifica ogni linguaggio nello spazio delle ipotesi dato inizialmente *indipendentemente dall'elenco*.

2. – Identificazione efficiente.

Nei problemi di identificazione sopra presentati, e ancora di più nelle applicazioni pratiche, emerge chiaramente il problema dell'*efficienza*, cioè della quantità di dati di cui il learner ha bisogno per formulare la congettura esatta, e del tempo di computazione necessario per emettere la congettura. È infatti chiaro che, mentre nella definizione formale e astratta di identificazione si richiede che essa avvenga dopo un numero finito, ma arbitrariamente grande, di passi, nella pratica il tempo necessario per identificare è di estrema importanza. Tale semplice osservazione costituisce la motivazione principale della tesi. Il problema dell'efficienza nell'identificazione non è nuovo ([5], [3]). L'interpretazione adottata nella tesi usa la parola «efficiente» come sinonimo di «polynomial-time» (p-time). Si considerano innanzitutto classi di funzioni (linguaggi) p-time computabili, ossia funzioni ricorsive calcolabili mediante una macchina di Turing deterministica che lavora in tempo polinomiale nella lunghezza dei dati (linguaggi aventi funzione caratteristica p-time). Indicheremo con P la classe di tali funzioni. Le idee fondamentali sono dunque le seguenti:

– Vengono dati a priori una classe $\mathcal{C}$ di oggetti (oggetti: funzioni o linguaggi p-time) ed una classe $\mathcal{R}$ di rappresentazioni (indici) per gli oggetti di $\mathcal{C}$. Ogni oggetto di $\mathcal{C}$ ha almeno un indice in $\mathcal{R}$.

– Viene fornito un elenco di esempi relativi ad un oggetto di $\mathcal{C}$: nel caso di

una funzione l'elenco dei valori da essa assunti, nel caso di un linguaggio l'elenco dei valori assunti dalla sua funzione caratteristica (*informant*).

– Un learner deve formulare delle congetture in $\mathcal{R}$ sull'oggetto incognito. Si richiede che il learner sia simulabile con una funzione p-time.

Il learner ha successo nell'identificare $\mathcal{C}$ se, per ogni $c \in \mathcal{C}$, le sue congetture si stabilizzano su un indice per c in $\mathcal{R}$ dopo che è stato fornito un numero di dati polinomiale nel minimo indice per c in $\mathcal{R}$. Per la precisione consideriamo due diversi paradigmi. Nel primo (*identificazione efficiente*) si richiede che il numero di dati necessario per raggiungere l'ipotesi esatta sia polinomiale nel *minimo indice* di c in $\mathcal{R}$. Nel secondo caso (*identificazione molto efficiente*) si richiede che tale numero di dati sia polinomiale nella *lunghezza del minimo indice* di c in $\mathcal{R}$. Alcuni aspetti fondamentali riguardanti l'identificazione efficiente sono i seguenti:

(1) Non si richiede l'identificazione dopo un numero fissato di passi, ma solo dopo un tempo non molto più lungo del più piccolo indice dell'oggetto da identificare. Quindi è possibile che il learner cambi parere dopo un numero molto grande di passi, ma solo se l'oggetto ha minimo indice molto grande. L'idea è che più l'oggetto da identificare ha una rappresentazione complessa, più tempo è concesso per l'identificazione.

(2) Si considera accettabile una identificazione se la congettura è nello spazio degli indici fissato a priori.

(3) Il successo del learner può dipendere non solo dalla classe da identificare, ma anche dalla classe di indici per essa adottata. Ad esempio, se ogni oggetto ha solo indici molto grandi, il learner ha più tempo a disposizione per identificare (il tempo dipende in modo polinomiale dal minimo indice).

(4) Sembra abbastanza ragionevole richiedere che il learner possa decidere rapidamente, dati un indice h di un oggetto e un sample S (insieme finito di esempi), se S è consistente con h o meno. Per questo motivo, ci siamo soffermati su quelle classi $\mathcal{C}$ di funzioni (linguaggi) in P e classi $\mathcal{R}$ di indici per $\mathcal{C}$ per cui il problema di computare (decidere), dati $h \in \mathcal{R}$ e $x \in \omega$, il valore $f(x)$ (se $x \in L$), ove $f(L)$ è la funzione (il linguaggio) in $\mathcal{C}$ rappresentata da h, sia in P. In tal caso diremo che $\mathcal{C}$ è *fortemente uniforme* (f.u.). Inoltre useremo il simbolo $\mathcal{R}_{\mathcal{C}}$ per denotare una classe di rappresentazioni per una classe f.u. $\mathcal{C}$ per cui vale la suddetta proprietà.

I risultati più importanti ottenuti sono una caratterizzazione per l'identificazione efficiente di classi f.u., attraverso cui si dimostra che l'intera classe P non è identificabile efficentemente. Nel caso molto efficiente diamo invece una condizione necessaria, che risulta anche sufficiente se e soltanto se $P = NP$. Esibiamo anche esempi di alcune classi f.u. note in matematica che risultano identificabili molto efficientemente, tra cui la classe dei polinomi a coefficienti interi positivi in una sola variabile e la classe delle funzioni resto modulo n.

L'identificazione efficiente di linguaggi su dati positivi risulta più difficile da trattare. Il motivo è che se nella classe ci sono due linguaggi L ed L' aventi un elemento x in comune, non possiamo identificare nessuno dei due in modo efficiente. Infatti su elenchi che iniziano con un numero arbitrariamente grande di «x», il learner non è in grado di decidere in modo rapido se il linguaggio rappresentato sia L o L'. A questo scopo, si è rivelata molto utile l'identificazione in mi-

sura. L'idea è che, dati una classe $\mathcal{C}$ f.u. e $L \in \mathcal{C}$, il learner riceve ad ogni passo n un elemento di L estratto casualmente in base ad una certa distribuzione di probabilità, in modo che, con probabilità 1, vengono generati tutti (e soli) gli elementi di L. Si richiede che il learner, dati un numero $\varepsilon > 0$ e $L \in \mathcal{C}$ avente minimo indice h in $\mathcal{R}_{\mathcal{C}}$, debba iniziare a formulare una congettura esatta in tempo polinomiale in (nella lunghezza di) h e $1/\varepsilon$, con probabilità $> 1 - \varepsilon$ (*identificazione in misura - molto - efficiente*). In altri termini il numero di passi concesso per raggiungere l'ipotesi esatta con probabilità $> 1 - \varepsilon$ può essere tanto più grande quanto più complicata è la descrizione del linguaggio e quanto più piccola è la probabilità dell'errore. Si dimostra che ogni classe di linguaggi $\mathcal{C}$ f.u. identificabile (molto) efficientemente su informant lo è anche in misura rispetto a particolari classi di distribuzioni di probabilità per $\mathcal{C}$.

Un approccio che invece affianca aspetti probabilistici a informazioni positive e negative per un linguaggio (informant) è presentato nella cosiddetta *identificazione probabilistica*. In tal caso, dati una classe $\mathcal{C}$ f.u. e $L \in \mathcal{C}$, il learner dispone ad ogni passo n del segmento iniziale di lunghezza $n + 1$ dell'informant per L e di una stringa binaria di lunghezza $n + 1$ che possiamo pensare generata con lanci casuali di moneta. Questo approccio vuole rappresentare l'idea che la congettura di un learner possa essere determinata, oltre che da informazioni «esatte» su L, anche da eventi esterni del tutto casuali ed indipendenti tra loro. Diciamo che il learner identifica probabilisticamente L in modo (molto) efficiente se, per ogni numero $\varepsilon > 0$, le congetture del learner in corrispondenza delle due sequenze generate si stabilizzano su indice per L in $\mathcal{R}_{\mathcal{C}}$ in tempo polinomiale nel (nella lunghezza del) minimo indice per L in $\mathcal{R}_{\mathcal{C}}$ con probabilità $> 1 - \varepsilon$. Si dimostra che, nel caso efficiente, l'identificazione probabilistica è equivalente a quella su informant (cioè i lanci di moneta non aiutano il learner), mentre nel caso molto efficiente si ha solo che l'identificazione su informant implica quella probabilistica.

BIBLIOGRAFIA

[1] D. ANGLUIN, *Inductive Inference of Formal Languages from Positive Data*, Information and Control, **45** (1980), 117-135.

[2] E.M. GOLD, *Language Identification in the Limit*, Information and Control, **10** (1967), 447-474.

[3] M.G. KEARNS and U.V. VAZIRANI, *An Introduction to Computational Learning Theory*, The MIT Press, Cambridge, Massachusetts, London, England, (1994).

[4] P. ODIFREDDI, *Classical Recursion Theory II*, in preparazione.

[5] D.N.OSHERSON, M. STOB and S. WEINSTEIN, *Systems that Learn. An Introduction to Learning Theory for Cognitive and Computer Scientists*, MIT Press, Cambridge MA (1986).

Dipartimento di Matematica, Università di Siena; e-mail: fontanis@unisi.it
Dottorato in Logica Matematica ed Informatica Teorica (sede amministrativa: Siena) - Ciclo IX
Direttore di ricerca: Prof. Franco Montagna