

Analisi della neutralità degli spazi di ricerca booleani in Programmazione Genetica

Riassunto

Yuri Pirola
matr. n. 060962

Relatore
Dott. Leonardo Vanneschi

Correlatore
Prof. Giancarlo Mauri

Anno accademico 2004/2005

Corso di Laurea Magistrale in Informatica
Dipartimento di Informatica, Sistemistica e Comunicazione - DISCo



Facoltà di Scienze Matematiche, Fisiche e Naturali
Università degli Studi Milano Bicocca



Obiettivi e motivazioni

Negli ultimi anni, le tecniche di Programmazione Genetica sono state in grado di risolvere con successo alcuni problemi reali considerati difficili. Tuttavia vi sono ancora alcune questioni irrisolte che ne impediscono l'utilizzo diffuso. Tra queste vi è la mancanza di un metodo per predire la difficoltà di un problema per la PG, cioè di predire se la PG sarà in grado di trovare soluzioni 'buone' al problema. La neutralità del paesaggio di fitness influisce sulle prestazioni della PG perché annulla il vantaggio rappresentato dal criterio di selezione. Di conseguenza si potrebbe ipotizzare la presenza di una relazione fra la difficoltà del problema e la neutralità del paesaggio ad esso associato. Il paesaggio di fitness è uno strumento teorico in grado di modellare, almeno in parte, il comportamento della PG, mentre la neutralità è il fenomeno di avere *differenti soluzioni* allo *stesso livello di qualità*. La tesi presenta uno studio della neutralità di una classe specifica di paesaggi di fitness. Confrontando paesaggi di fitness indotti da problemi differenti, si cercherà di trovare alcune differenze che potrebbero giustificare il diverso grado di difficoltà che la PG incontra nel risolvere il problema.

Questo studio differisce da contributi precedenti perché (1) considera la PG standard, cioè quella basata su rappresentazioni ad albero, e perché (2) lo studio condotto sul paesaggio è avvenuto senza modificarlo.

I principali contributi originali sono: (a) lo studio teorico delle più importanti caratteristiche di alcuni paesaggi di fitness, (b) la definizione di alcune nuove misure che aiutano nel caratterizzare la neutralità del paesaggio e (c) l'analisi sperimentale di alcuni paesaggi di fitness.

Innanzitutto sono stati considerati alcuni paesaggi di taglia 'ridotta' mentre, in seguito, è stata effettuata un'analisi campionaria di paesaggi più 'grandi'. Le tecniche di campionamento generalmente utilizzate in letteratura non sono adatte per la tipologia di analisi che ci si propone. Esse infatti non considerano alcune importanti proprietà del paesaggio che si è interessati ad analizzare. Di conseguenza è stata proposta ed utilizzata una nuova metodologia di campionamento del paesaggio di fitness.

I risultati sperimentali sostengono l'adeguatezza, per la classe di paesaggi che si è considerato, della metodologia di campionamento presentata qui per la prima volta. Inoltre, le differenti caratteristiche del paesaggio che si sono evidenziate sembrano giustificare i differenti gradi di difficoltà dei problemi associati.

Programmazione Genetica e Paesaggi di Fitness

La Programmazione Genetica è una tecnica di apprendimento automatico introdotta da Koza nel 1992 [1] per generare programmi in grado di eseguire

un compito specifico. La PG si ispira alla teoria darwiniana sull'evoluzione delle specie e opera selezionando e variando una *popolazione* di programmi. Il processo di selezione è guidato da un criterio di qualità espresso tramite una funzione di *fitness*. La variazione, invece, è compiuta tramite un insieme di operatori genetici che trasformano uno o più individui, chiamati genitori, in altrettanti nuovi individui, chiamati figli.

La letteratura associata alla PG ha proposto differenti schemi di rappresentazione degli individui. Il più utilizzato è quello dovuto a Koza che si basa su una rappresentazione delle soluzioni tramite strutture ad albero. Esistono, comunque, altre varianti basate su strutture lineari o su grafi.

Un *paesaggio di fitness* può essere definito come una tripla $\mathcal{P} = (\mathcal{S}, \mathcal{V}, f)$ dove \mathcal{S} è l'insieme di tutte le soluzioni candidate, $\mathcal{V} : \mathcal{S} \rightarrow 2^{\mathcal{S}}$ è la funzione di vicinato che specifica, per ciascuna soluzione $s \in \mathcal{S}$, l'insieme dei suoi vicini $\mathcal{V}(s)$, mentre $f : \mathcal{S} \rightarrow \mathbb{R}$ è la funzione di fitness. La relazione di vicinato è generalmente definita in funzione degli operatori di variazione utilizzati, cioè \mathcal{V} può essere definita come $\mathcal{V}(s) = \{s' \in \mathcal{S} | s' \text{ può essere ottenuta da } s \text{ con una singola variazione}\}$.

In alcuni casi, anche se la dimensione dello spazio di ricerca è notevole, f può assumere solo un insieme limitato di valori. Di conseguenza, molte soluzioni condividono la stessa fitness e si può dire che il paesaggio ha un alto grado di neutralità [2]. Data una soluzione s , si può definire un particolare sottoinsieme di $\mathcal{V}(s)$: quello composto dai vicini 'neutri'. Formalmente, il *vicinato neutro* di s è l'insieme $\mathcal{N}(s) = \{s' \in \mathcal{V}(s) | f(s') = f(s)\}$.

Date queste definizioni, è possibile immaginare un paesaggio di fitness come composto da un insieme (potenzialmente numeroso) di *zone di neutralità*. Più formalmente possiamo definire il concetto di *rete di neutralità* [3] come una componente connessa del grafo $(\mathcal{S}, E_{\mathcal{N}})$ dove $E_{\mathcal{N}} = \{(s_1, s_2) \in \mathcal{S}^2 | s_2 \in \mathcal{N}(s_1)\}$.

Il problema dell'Even-Parity

Il problema dell'even-k parity [1] consiste nel determinare un'espressione booleana di k variabili che vale Vero se e solo se sono vere un numero pari di esse. La fitness è calcolata come il numero degli errori diviso il numero di combinazioni possibili degli ingressi (cioè 2^k).

Il problema dell'even-parity è conosciuto ed utilizzato in letteratura perché si ritiene che la funzione da approssimare è difficile da ottenere tramite molte tecniche di Apprendimento Automatico.

L'insieme di tutte le soluzioni candidate è l'insieme delle espressioni booleane ben formate rispetto a un insieme di operatori \mathcal{F} e un insieme di variabili \mathcal{T} . Per limitare la cardinalità di \mathcal{S} si è imposto un vincolo sull'altezza

massima dell'albero che rappresenta l'espressione. L'insieme \mathcal{T} è composto da k variabili booleane mentre si è scelto di studiare due differenti insiemi di operatori: $\{\text{XOR}; \text{NOT}\}$ e $\{\text{NAND}\}$. Questi due insiemi di operatori inducono due differenti paesaggi di fitness (denotati da $\mathcal{P}_{(k,h)}^{\{\text{XOR}; \text{NOT}\}}$ e $\mathcal{P}_{(k,h)}^{\{\text{NAND}\}}$, dove k è l'ordine del problema e h è l'altezza massima consentita) con differenti gradi di difficoltà: quello indotto da $\{\text{XOR}; \text{NOT}\}$ è più 'facile' da esplorare rispetto a quello indotto da $\{\text{NAND}\}$.

La definizione della struttura di vicinato è subordinata alla scelta degli operatori di variazione. Si è proposto una versione ridotta degli operatori di mutazione strutturale introdotti in [4] a cui ci riferiremo con il nome di *mutazioni strutturali strette* per distinguerli dai precedenti. Questo insieme di mutazioni è sufficientemente semplice da studiare ma, al contempo, fornisce sufficiente capacità esplorativa alla PG.

Il paesaggio $\mathcal{P}^{\{\text{XOR}; \text{NOT}\}}$ presenta alcune importanti proprietà che sono state derivate e dimostrate per la prima volta nella tesi. Innanzitutto, indipendentemente dall'insieme degli operatori che si considera, se un'espressione non contiene almeno un'occorrenza di ciascuna variabile, allora la sua fitness è esattamente uguale a 0.5. Per questo motivo, una vasta maggioranza di individui dei paesaggi di fitness dell'even-parity ha fitness pari a 0.5. Inoltre, un'espressione di $\mathcal{P}^{\{\text{XOR}; \text{NOT}\}}$ può assumere solo tre distinti valori di fitness, ovvero 0, 0.5 e 1.

La scelta degli operatori di mutazione strutturale stretta ha permesso di definire altre proprietà del paesaggio $\mathcal{P}^{\{\text{XOR}; \text{NOT}\}}$: (a) esiste *solo* una rete di neutralità a fitness 0.5 (che chiameremo *rete centrale*), (b) tutte le altre reti di neutralità sono composte da un solo individuo (chiameremo queste reti col nome di *reti periferiche*) e (c) tutte le reti periferiche sono connesse con la rete centrale tramite una mutazione. La dimostrazione di queste proprietà è discussa per esteso nella tesi.

La caratterizzazione teorica del paesaggio $\mathcal{P}^{\{\text{XOR}; \text{NOT}\}}$ sarà di grande importanza, come spiegato in seguito, per la valutazione degli effetti dovuti al campionamento.

Campionamento dei paesaggi di fitness

Il campionamento dei paesaggi di fitness è, spesso, un'attività necessaria perché il numero di soluzioni che li compongono è molto elevato. Nel corso della tesi è stata derivata un'equazione di ricorrenza che 'conta' il numero di soluzioni degli spazi di ricerca booleani. Essa mostra che la dimensione dello spazio di ricerca cresce molto velocemente (più velocemente di α^x) al crescere dell'altezza massima ammessa. Inoltre, anche l'insieme degli operatori e delle

variabili che si utilizza incide notevolmente sulla dimensione dello spazio di ricerca.

La generazione di campioni adatti per l'analisi dei paesaggi di fitness booleani (e in particolare relativi all'even-parity) è un compito difficile. Le tecniche di campionamento tradizionali non offrono una 'vista' significativa del paesaggio di fitness perché spesso ne ignorano alcune parti importanti (come quelle composte da individui di fitness diversa da 0.5) oppure non ne 'rispettano' la struttura reale (ad es. generano molti individui isolati). Di conseguenza, è stato proposto un nuovo processo di campionamento che combina tecniche consolidate con altre fasi progettate per riprodurre la struttura 'locale' del paesaggio. L'obiettivo di questo processo è la generazione di campioni contenenti individui di molti (possibilmente tutti i possibili) livelli di fitness e che formano, tra loro, un insieme connesso di reti di neutralità sufficientemente 'grandi'. Il processo di campionamento è composto da tre fasi chiamate *campionamento di Metropolis modificato*, *espansione verticale* e *espansione orizzontale*. Il campionamento di Metropolis modificato genera un campione C di soluzioni con valori di fitness che 'coprono' possibilmente tutto l'intervallo ammissibile. L'espansione verticale 'arricchisce' il campione C con soluzioni appartenenti al vicinato (preferibilmente *non neutro*) di individui del campione. L'espansione orizzontale, infine, cerca di 'allargare' le reti di neutralità *incomplete*¹ di taglia inferiore a un certo limite prefissato. La generazione casuale di individui è stato un altro tema trattato nel corso della tesi ed ha portato alla definizione di una variante del tradizionale metodo GROW adatta a questa tipologia di individui.

Analisi sperimentale e sviluppi futuri

Nel corso di questo lavoro sono state compiute due differenti tipologie di analisi: la prima ha considerato paesaggi di fitness di taglia ridotta, in modo da essere in grado di generare tutti gli individui che li compongono, mentre la seconda ha studiato campioni (ottenuti tramite la nostra tecnica di campionamento) di paesaggi di taglia maggiore. L'analisi esaustiva ha considerato i paesaggi $\mathcal{P}_{(2,3)}^{\{XOR; NOT\}}$ e $\mathcal{P}_{(2,3)}^{\{NAND\}}$, cioè i due paesaggi associati all'even-2 parity e con altezza massima delle soluzioni uguale a 3. La prima analisi campionaria ha considerato ancora il problema dell'even-2 parity ma con alberi di altezza massima maggiore (nello specifico uguale a 7). Il confronto di questi risultati con quelli ottenuti precedentemente e con la caratterizzazione teorica di $\mathcal{P}^{\{XOR; NOT\}}$ ha permesso di valutare empiricamente la validità del metodo di campionamento proposto. A seguito di questo confronto, la tecnica di

¹ Una rete di neutralità è detta *incompleta* se non tutti i suoi individui appartengono al campione.

campionamento proposta sembra ‘comportarsi bene’ su questa tipologia di paesaggi. Il passo successivo è stato lo studio di campioni dei paesaggi associati all’even-4 parity. Per garantire la presenza di almeno un ottimo, anche l’altezza massima delle soluzioni è stata incrementata a 8.

Tutte le analisi hanno considerato lo stesso insieme di *misure di rete*, cioè misure calcolate su reti di neutralità. Questo insieme include alcune misure ‘tradizionali’ (come la taglia della rete, la sua fitness, ecc.) ed altre nuove misure relative alla neutralità che sono state specificatamente definite. In particolare si è definito: (1) il *tasso medio di neutralità di rete*, che quantifica il numero di mutazioni neutre tra i suoi individui, (2) la *Δ -fitness media di rete*, che quantifica il *guadagno medio* di fitness che si ottiene a seguito di una mutazione di un individuo della rete, (3) il *tasso di subottimi*, che quantifica il numero di soluzioni di una rete che non possono generare figli ‘migliori’ e (4) il *tasso di subpessimi*, che quantifica il numero di soluzioni di una rete che non possono generare figli ‘peggiori’.

Studiando il tasso medio di neutralità, si è osservato che le reti del paesaggio $\mathcal{P}^{\{\text{NAND}\}}$ con valori di fitness alti sembrano essere maggiormente neutre rispetto a reti con fitness buona. Il paesaggio $\mathcal{P}^{\{\text{XOR}; \text{NOT}\}}$ presenta un’unica rete di taglia maggiore di 1, quindi il tasso medio di neutralità delle reti periferiche (che hanno fitness sia buona che ‘cattiva’) è pari a 0.

L’analisi delle altre tre misure ha evidenziato che reti di $\mathcal{P}^{\{\text{NAND}\}}$ con fitness buona, in un certo senso, ‘resistono’ al miglioramento, ovvero le mutazioni che hanno origine in queste reti mostrano la tendenza a peggiorare (o migliorare di poco) la fitness. In $\mathcal{P}^{\{\text{XOR}; \text{NOT}\}}$, invece, sono permessi ampi miglioramenti (o peggioramenti) di fitness: la ricerca dell’ottimo, di conseguenza, sembra essere facilitata.

Questi risultati possono parzialmente giustificare il motivo per il quale il problema dell’even-parity è più facile per la PG se si utilizzano gli operatori $\{\text{XOR}; \text{NOT}\}$ invece di $\{\text{NAND}\}$. Questi risultati valgono sia per gli spazi ‘limitati’ che sono stati studiati in via esaustiva sia per gli spazi di taglia maggiore, studiati per via campionata.

Le tecniche presentate durante il lavoro di tesi sono di carattere generale e, dunque, possono essere applicate ad ogni tipologia di spazio di ricerca della PG. Gli sviluppi futuri di questa tesi includono, quindi, l’analisi di altri problemi, sia relativi alla regressione simbolica di espressioni booleane (come è stato per l’even-parity) sia di nuove tipologie. Da questi ulteriori studi potrebbero scaturire, inoltre, nuove misure della difficoltà del problema basate sulla neutralità del paesaggio. Un’altra direzione di sviluppo, infine, è rappresentata dallo studio più approfondito della tecnica di campionamento proposta, estendendone l’analisi sperimentale su nuove tipologie di paesaggio oppure formalizzandola attraverso strumenti statistico–matematici.

Riferimenti bibliografici

- [1] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.
- [2] C. M. Reidys and P. F. Stadler. Neutrality in fitness landscapes. *Applied Mathematics and Computation*, 117(2-3):321-350, 2001.
- [3] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. In *Proc. R. Soc. London B.*, volume 255, pages 279-284, 1994.
- [4] L. Vanneschi, M. Tomassini, P. Collard, and M. Clergue. Fitness distance correlation in structural mutation genetic programming. In Ryan Conor *et al.*, editor, *Genetic Programming, Proceedings of EuroGP'2003*, volume 2610 of *LNCS*, pages 455-464, Essex, 14-16 Apr. 2003. Springer-Verlag.