UNA PROCEDURA DI TAGLIO A GRADINI PER L'ESTRAZIONE DELLA PARTIZIONE OTTIMALE DA UN DENDROGRAMMA

Dario Bruzzese[1], Umberto Giani[1], Domenico Vistocco[2]
1 Dipartimento di Scienze Mediche Preventive, Università di Napoli "Federico II",
2 Dipartimento di Scienze Economiche, Università di Cassino

Introduzione

L'individuazione, mediante procedure di classificazione, di tipologie omogenee è tutt'oggi un obiettivo di primaria importanza in molti contesti applicativi, dall'analisi di profili di espressione genica, allo studio delle reti dinamiche biologiche. Tuttavia la mancanza di informazioni a priori rende spesso necessario l'impiego di tecniche non supervisionate (tecniche di clustering in senso stretto) in cui il problema della scelta del numero di classi è di cruciale importanza.

Obiettivi

Il presente lavoro introduce una procedura automatica per la individuazione della partizione ottimale a partire dal dendrogramma prodotto da una Classificazione Gerarchica. La soluzione si ottiene attraverso un opportuno test di permutazione applicato iterativamente sui differenti livelli di aggregazione dell'albero. La partizione così individuata è costituita da gruppi che non necessariamente rispettano il vincolo della classificabilità, imposto invece dal taglio orizzontale di un dendrogramma.

Metodi

L'algoritmo opera sulla struttura ad albero di un dendrogramma prodotto da una classificazione gerarchica seguendo però un approccio divisivo dal momento che percorre l'albero dal nodo radice sino al raggiungimento della regola di arresto.

Ai fini della descrizione dell'algoritmo, si indichi con n il numero di oggetti da classificare, con k lo step in corrispondenza del quale due generiche classi C^k_l, C^k_r sono state aggregate $k=1,\ldots,n-1$ (dove $k=1$ indica il livello del nodo radice), con $h(C^k_l, C^k_r)$ l'altezza in corrispondenza della quale le due classi sono state aggregate e infine con $h(C^k_j)$ l'altezza in corrispondenza della quale si è costituita la classe C^k_j, $j \in \{l,r\}$, $k=1,\ldots n-1$ (nel caso di classi singleton l'altezza è posta pari a 0). Per ogni k e quindi per ogni coppia C^k_l, C^k_r, viene realizzato un test di permutazione allo scopo di sottoporre a verifica l'ipotesi H_0: $C^k_l \equiv C^k_r$ e cioè l'ipotesi che i due insiemi provengono dalla stessa *popolazione*. Sotto tale ipotesi, la riallocazione casuale, in due nuove classi, delle unità statistiche che popolano C^k_l, C^k_r, non dovrebbe alterare la struttura di aggregazione che le ha prodotte. Siano allora ${}_mC^k_l$, ${}_mC^k_r$ le due classi ottenute a seguito della procedura di permutazione e rispettando il vincolo della numerosità.

Per ognuna di queste, un nuovo dendrogramma è generato; siano $h({}_mC^1_l)$ e $h({}_mC^1_r)$ le altezze in corrispondenza delle quali le due classi ${}_mC^k_l$, ${}_mC^k_r$ sono ricostituite (quindi le altezza dei nodi radice dei rispettivi dendrogrammi).

Il test si basa sul confronto tra il costo minimo necessario all'aggregazione delle due classi originali e quello associato alle classi permutate. Il primo è definito come:

$$cost(C^k_l, C^k_r) = \frac{\max_{j\in\{l,r\}}\left(h\left(C^k_j\right)\right) - \min_{j\in\{l,r\}}\left(h\left(C^k_j\right)\right)}{h\left(C^k_l, C^k_r\right) - \max_{j\in\{l,r\}}\left(h\left(C^k_j\right)\right)}$$

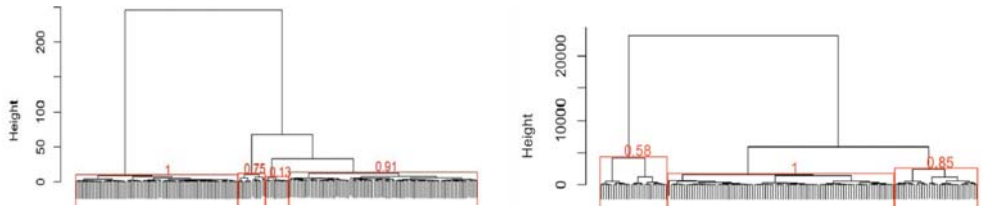
Quello associato alle due classi permutate, $cost({}_mC^k_l, {}_mC^k_r)$, sostituisce ai valori $h(C^k_j)$ i corrispondenti valori $h({}_mC^k_j)$, j ∈ {l,r}. La procedura di permutazione ora descritta è ripetuta M volte e il pvalue Montecarlo è calcolato come (Good, 2005):

$$p = \frac{\#\left\{cost\left({}_mC^k_l, {}_mC^k_r\right) \le cost\left(C^k_l, C^k_r\right)\right\}}{M}$$

La procedura si arresta quando, per ogni sottoalbero, si incontra quel livello di aggregazione in corrispondenza del quale l'ipotesi H_0 non può essere rifiutata.

Risultati

La procedura è stata applicata su differenti insiemi di dati; per ragioni di brevità solo due casi saranno discussi. In entrambi, il dendrogramma è stato generato utilizzando la distanza Euclidea ed il criterio di aggregazione di Ward. Nella procedura di taglio sono state invece utilizzate 999 permutazioni per ogni ciclo Montecarlo ed un livello di significatività pari a 0.05. In figura *1a* è riportato il dendrogramma generato su un sottoinsieme di 205 profili di espressione genica contenuti nel dataset *Yeast Galactose* (Ideker et alt. 2001) e che riflettono 4 differenti categorie funzionali della Gene Ontology. Come si evince dal grafico, la procedura proposta ha consentito di estrarre correttamente le 4 differenti classi prodotte dal dendrogramma. La figura *1b* si riferisce invece al dataset *Diabetes* (Reaven e Miller, 1979) che descrive 145 osservazioni appartenenti a tre differenti tipologie cliniche (soggetti sani, diabetici subclinici e diabetici clinici). Anche in questo caso il metodo ha estratto la corretta partizione dall'albero di classificazione gerarchico.



(a) (b)

Fig. 1: Risultati della procedura di taglio applicata al dataset *Yeast Galactose (a)* e al dataset *Diabetes* (b). In rosso sono evidenziate le classi estratte dall'algoritmo ed i rispettivi pvalue.

Conclusioni

La procedura descritta individua automaticamente il taglio di un albero di classificazione gerarchica. La soluzione risulta, alla luce delle simulazioni effettuate, stabile rispetto al numero di cicli montecarlo; la scelta del livello di significatività consente invece di modulare il livello di granularità della partizione. Rispetto ad altri criteri tradizionali ha il vantaggio di non limitare la ricerca ad una sola classe di partizioni ma consente di scegliere nell'insieme di tutte le partizioni possibili.

Bibliografia

1. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner RE, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systemically perturbed metabolic network. *Science* 2001, 292:929-934.
2. G.M. Reaven and R.G. Miller, *Diabetologica* 1979, 16:17-24
3. P. Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses* 2005, Springer, Germany