UN'INTRODUZIONE ALLA METODOLOGIA CART

Giuseppe Espa
Istituto di Statistica e Ricerca Operativa
Università degli Studi di Trento

1. Rappresentazione della struttura dei dati per mezzo di alberi

CART è un acronimo di ***C**lassification **A**nd **R**egression **T**rees*, una nuova procedura statistica introdotta da Leo Breiman, Jerom Friedman, Richard Olshen e Charles Stone nel 1984. Come il nome stesso suggerisce, il **CART** è una singola procedura che può essere utilizzata per analizzare dati qualitativi (classificazione) o quantitativi (regressione) per mezzo della stessa metodologia. Una caratteristica definitoria del **CART** è che tale tecnica presenta i suoi risultati nella forma di alberi di decisione, cosa che si discosta in modo rilevante dal comportamento delle procedure statistiche più tradizionali. La struttura ad albero dell'*output* consente al **CART** di trattare strutture di dati molto complesse, producendo in uscita dei grafici facilmente interpretabili. Nello spirito dell'analisi esplorativa dei dati, il **CART** è una metodologia non parametrica che dialoga con l'utente per mezzo di figure.

Come esempio si consideri l'albero riportato in Figura 1.1 (tratta da Breiman *et al.*, 1984) costruito per analizzare dati relativi a pazienti colpiti da infarto e ricoverati presso il San Diego Medical Center dell'Università della California.

Lo *staff* medico del reparto di cardiologia dell'ospedale ha necessità di determinare con la massima accuratezza e nel minor tempo possibile se ciascun paziente ricoverato è a rischio di un secondo e letale infarto o meno. I pazienti ad alto rischio (AR) vengono trasferiti presso un'unità coronarica per un monitoraggio continuo e costante mentre i pazienti a basso rischio (BR) vengono avviati alla degenza presso reparti ordinari. La scelta non è indifferente dal momento che le possibilità di ricovero nel reparto di terapia intensiva sono molto limitate ed hanno un costo sostenuto.

I dati per l'analisi consistono di 19 variabili rilevate, nelle prime 24 ore che seguono il ricovero, su 215 pazienti. I dati includono variabili quali:

i) pressione sanguigna (PS);
ii) età (ET);
iii) presenza di tachicardia (T);
iv) concentrazione di certi enzimi nel sangue;
v) etc..

Nell'arco dei 30 giorni che seguono il ricovero 37 pazienti muoiono (il gruppo AR) mentre ne sopravvivono 178 (il gruppo BR).

Analizzando in modo sistematico questi dati il **CART** produce lo schema riportato in Figura 1.1.

Nella Figura 1.1 si distinguono due tipologie di nodo: i nodi non terminali, i quali hanno due discendenti diretti, ed i nodi terminali (che riportano in basso l'etichetta assegnata loro) i quali non subiscono ulteriori bipartizioni. Si noti che la decisione di bipartire è basata su una singola variabile e che sono state utilizzate soltanto 3 delle 19 variabili rilevate sui pazienti.

Il semplice diagramma riportato in Figura 1.1 costituisce l'essenza di un'analisi di classificazione condotta mediante il **CART**. Il processo comincia a partire da un nodo radice e, attraverso una serie di domande che contemplano solo risposte SI/NO, vengono generati i nodi discendenti. Ogni *split*, usualmente basato sul valore di una singola variabile, divide sempre un nodo genitore in due nodi figli.

L'esempio riportato costituisce evidentemente un problema a due classi ma in linea teorica non esiste limite al numero di classi che il **CART** può trattare. Semmai limitazioni al riguardo provengono dalla dotazione *hardware* di cui si dispone.

Se la variabile dipendente fosse stata quantitativa, il **CART** avrebbe generato una sequenza analoga di nodi i quali avrebbero però contenuto in luogo delle classi dei valori previsti.

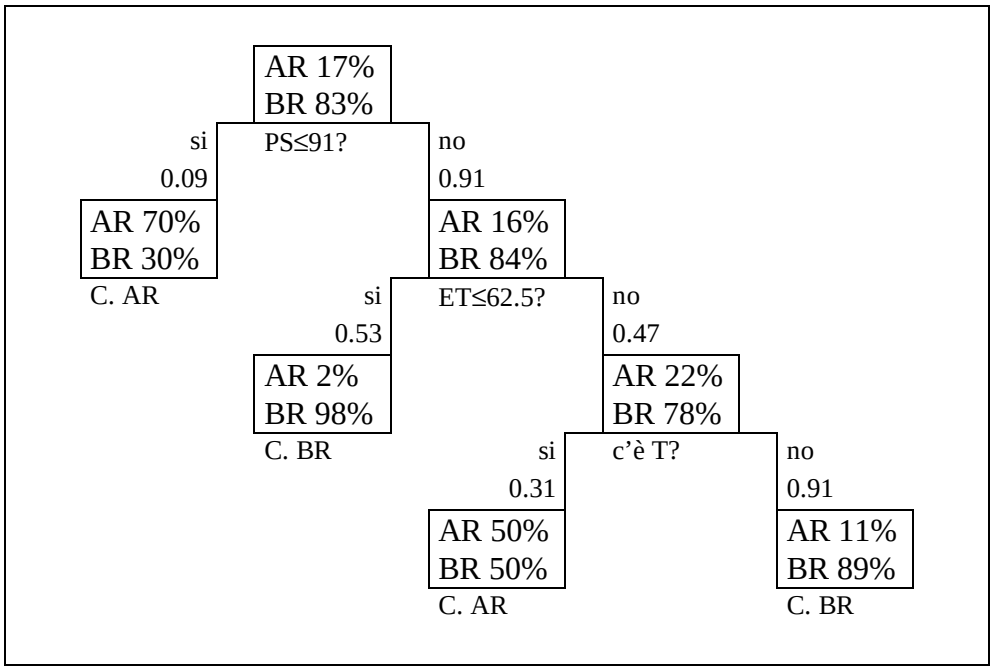


Figura 1.1 Esempio di struttura ad albero.

Il fascino del **CART** dovrebbe a questo punto essere evidente: l'analisi viene condotta principalmente in modo automatico ed i risultati che essa produce sono coincisi e facili da comprendere.

Il **CART** è stato applicato nell'ambito di numerose discipline fra le quali spiccano l'ingegneria, la diagnosi medica, l'economia, la sociologia e l'analisi di mercato. In particolare nella letteratura specialistica si segnalano studi sulla segmentazione dei mercati, sui rischi connessi al credito, sul controllo statistico della qualità, sul grado di diffusione di un tumore, sulla classificazione di cellule, sulla mortalità infantile, sull'inquinamento dell'aria, sulla classificazione di immagini digitali.

2. Metodologia CART: qualche intuizione

La metodologia **CART** è un esempio di ciò che tecnicamente prende il nome di *partizionamento binario ricursivo*. Il processo è *binario* in quanto un nodo genitore è bipartito sempre in esattamente due nodi figli ed è *ricursivo* perché la bipartizione può essere ripetuta trattando ogni nodo figlio come un nodo padre.

Gli elementi chiave di un'analisi **CART** sono un insieme di regole per:

i) *splittare* ogni nodo dell'albero;
ii) stabilire quando un albero è completo;
iii) assegnare una etichetta (od un valore previsto nel caso della regressione) ad ogni nodo terminale.

Qui di seguito verranno fornite alcune intuizioni circa tali ingredienti fondamentali.

2.1 Regole di bipartizione

Lo *split* sul nodo radice dell'albero riportato in Figura 1.1 poteva anche essere effettuato con riferimento ad altre regole di bipartizione. Ad esempio le domande cui fornire una risposta SI/NO potevano essere le seguenti:

L'età è ≤55? oppure *La pressione sanguigna è ≤ 102?*

A tali regole di bipartizione corrispondono gli *split* del nodo radice riportati rispettivamente in Figura 2.1 e 2.2

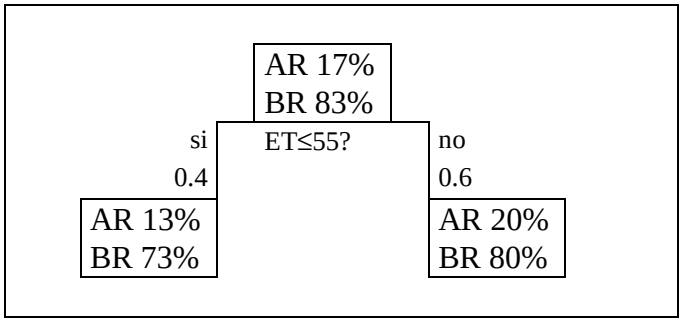


Figura 2.1 *Splitting* del nodo radice con la regola di bipartizione *L'età è ≤55?*.

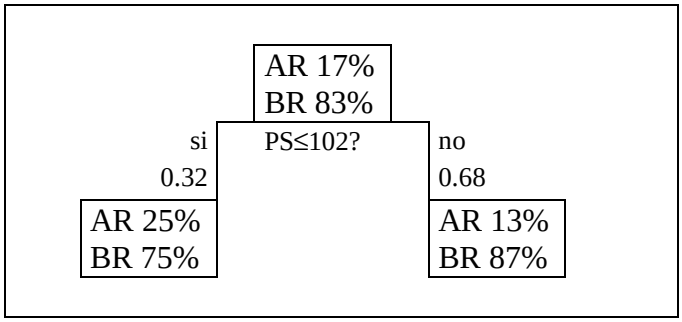


Figura 2.2 *Splitting* del nodo radice con la regola di *La pressione sanguigna è ≤ 102?*.

La metodologia **CART** si articola guardando a tutti i possibili *split* per tutte le variabili considerate nell'analisi. Nel *data-set* utilizzato per costruire l'esempio riportato nel precedente paragrafo possono essere rilevati al massimo 215 differenti valori per ognuna delle 19 variabili rilevate. Pertanto il **CART** per bipartire il nodo radice deve considerare al massimo 215×19=4085 possibili *split*. Qualsiasi problema pratico sarà

costituito da un numero finito di *split* candidati ed il **CART** condurrà una bruta ricerca all'interno di questo insieme. A questo punto è spontaneo domandarsi quale *split* scegliere fra quelli analizzati. Un criterio è quello di ordinare tutti gli *split* candidati in base ad un criterio di bontà di bipartizione e scegliere quello che produce la migliore *performance*.

2.2 Arresto del processo ed assegnazione della classe

Dopo che si è individuato lo *split* migliore per il nodo radice, il **CART** ripete il processo di ricerca per ogni nodo figlio continuando a bipartire finché non è più possibile splittare o finché il processo non è arrestato per qualche ragione (ad esempio perché un nodo è costituito da un numero troppo esiguo di casi). Il processo di bipartizione è impossibile quando un nodo è costituito da un solo caso oppure quando tutti i casi che compongono un nodo afferiscono alla stessa classe.

Quando un nodo viene riconosciuto come terminale bisogna stabilire come classificare i casi che in esso sono contenuti. Un semplice criterio al riguardo consiste nell'invocare una regola di pluralità: al nodo si assegnerà l'etichetta del gruppo maggiormente rappresentato.

Storicamente la costruzione della prima generazione di alberi di classificazione avveniva *splittando* ogni nodo finché non smetteva di essere soddisfatto un qualche criterio di bontà di bipartizione. Quando la qualità di un particolare nodo cadeva sotto una certa soglia, il nodo raggiunto veniva dichiarato terminale. Quando tutti i rami uscenti dal nodo radice raggiungevano dei nodi terminali, il processo di costruzione dell'albero veniva considerato ultimato. Sebbene ormai datata questa tecnica è incorporata in molti *software* commerciali ma purtroppo conduce a valutazioni spesso erronee. La filosofia del **CART** è completamente diversa. Invece di cercare di stabilire se un nodo è terminale o meno, il **CART** procede diramando l'albero finché è possibile ed esaminando sotto-alberi ottenuti dalla potatura dei rami dell'albero di partenza. Tra tutti i sotto-alberi possibili si sceglierà il migliore in base ad un criterio detto del *minimo costo-complessità*.

3. Alcuni commenti statistici

Alcuni aspetti tecnici della metodologia **CART** sono di primario interesse per gli statistici. Qui di seguito verranno elencati i principali sottolineando, laddove ce ne sia necessità ed in linguaggio non tecnico, l'idea che ne è alla base.

*Il **CART** è una procedura non parametrica.*

*Il **CART** non richiede che le variabili siano selezionate in anticipo.*

I risultati sono invarianti rispetto a trasformazioni monotone delle variabili indipendenti.
Ciò comporta che non c'è alcuna necessità in ambito applicativo di sperimentare trasformazioni monotone delle variabili indipendenti (logaritmi, radici quadrate, elevamento a potenza positiva, etc.). Nel **CART** tali variazioni non modificano i risultati a meno che lo *split* sia basato su combinazioni lineari di variabili.

*Il **CART** può utilizzare data-set con struttura complessa.*
Al contrario dei modelli parametrici che sono stati messi a punto allo scopo di mettere in luce singole strutture dominanti nei dati, il **CART** è designato alla scopo di trattare dati dotati di struttura multipla.

*Il **CART** è estremamente robusto all'effetto degli outliers.*

*Il **CART** può utilizzare combinazioni lineari delle variabili per determinare gli split.*

*Il **CART** può utilizzare congiuntamente variabili categoriche e continue.*
In conseguenza di ciò il **CART** non richiede alcun pre-processamento dei dati. In particolare non è necessario discretizzare le variabili quantitative.

*Il **CART** può mettere in luce dipendenze ed interazioni.*

*Il **CART** può processare casi con dati mancanti.*

I principali vantaggi che derivano dall'uso del **CART** possono essere schematicamente individuati nei seguenti punti:

*Il **CART** produce alberi ottimali.*

*Il **CART** utilizza strumenti molto sofisticati per stabilire la sua accuratezza.*

*Il **CART** può utilizzare la stessa variabile in punti differenti dell'albero.*

*Il **CART** può essere utilizzato a supporto dei modelli parametrici convenzionali.*

Infine il **CART** può essere utilmente impiegato per:

l'analisi esplorativa dei dati;

la selezione di variabili in ambito modellistico tradizionale;

l'individuazione di interazioni rilevanti;

il miglioramento di modelli;

la semplificazione di processi di decisione;

la selezione di predittori;

la conversione di variabili quantitative in variabili qualitative;

il trattamento di dati mancanti.

4. Riferimenti bibliografici

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984) *Classification and Regression Trees*, Wadsworth International Group, Belmont, California.