Capitolo VII

TECNICHE DI SEGMENTAZIONE GERARCHICA

di Luigi Grossi

1. *Introduzione*

Le tecniche di segmentazione vengono utilizzate per individuare l'appartenenza di unità statistiche alle classi d'una variabile dipendente conoscendo i valori o le modalità d'una o più variabili esplicative. La regola individuata viene successivamente impiegata per classificare nuove unità statistiche di cui si ignora la categoria d'appartenenza. L'utilizzazione degli algoritmi di segmentazione rientra nell'ambito delle procedure esplorative dei dati (1). Esse possono essere convenientemente utilizzate qualora gli assunti teorici e distributivi dei metodi di classificazione classici (analisi discriminante, modelli log-lineari) non risultino sostenibili. I risultati delle tecniche di segmentazione vengono solitamente visualizzati attraverso strutture grafiche gerarchiche dette «alberi».

L'output grafico della segmentazione presenta punti di contatto con il dendrogramma della *cluster analysis*. Infatti, nello stadio finale entrambe le procedure producono una partizione delle unità statistiche. Nonostante ciò le differenze sono sostanziali. L'applicazione della segmentazione richiede la conoscenza a priori della classe di appartenenza delle unità. Scopo della *cluster analysis* è invece quello di costruire gruppi di unità statistiche partendo da un insieme indistinto. Inoltre, la segmentazione viene operata utilizzando una sola variabile (selezionata fra tutte le variabili a disposizione) ad ogni passo, mentre la formazione dei gruppi nella *cluster analysis* viene effettuata in base al calcolo di misure di distanze fra le unità statistiche calcolate utilizzando tutte le variabili a disposizione. Infine, la regola di classificazione

---

(1) Le tecniche di segmentazione gerarchica si sono rivelate particolarmente efficaci nella individuazione di strutture latenti in *data set* molto numerosi. Per tale motivo vengono spesso annoverate fra le tecniche di *data mining*.

individuata attraverso gli algoritmi di segmentazione, viene utilizzata per prevedere la collocazione di unità statistiche di cui non si conosce la classe di appartenenza.

Le tecniche di segmentazione in ambito economico-aziendale hanno interessanti applicazioni quali il *credit scoring* (2) (Hand e Henley, 1997), la previsione dei fallimenti (Grossi e Ganugi, 1999) e delle insolvenze (Centrale dei Bilanci, 1998), la segmentazione dei mercati (Molteni, 1993).

A titolo d'esempio, si prendano in considerazione i dati relativi ad un'analisi di *credit scoring* riassunti nella tab. 7.1 (3).

TAB. 7.1. *Classificazione di 323 clienti d'un istituto di credito secondo la regolarità nella restituzione del prestito.*

				CREDIT RANKING		Totale
				Affidabile	Non affidabile	
PAGA	sett.	ETA'	<=35	15	143	158
			>35	7	0	7
	mens.	ETA'	<=25	25	24	49
			>25	108	1	109
Totale				155	168	323

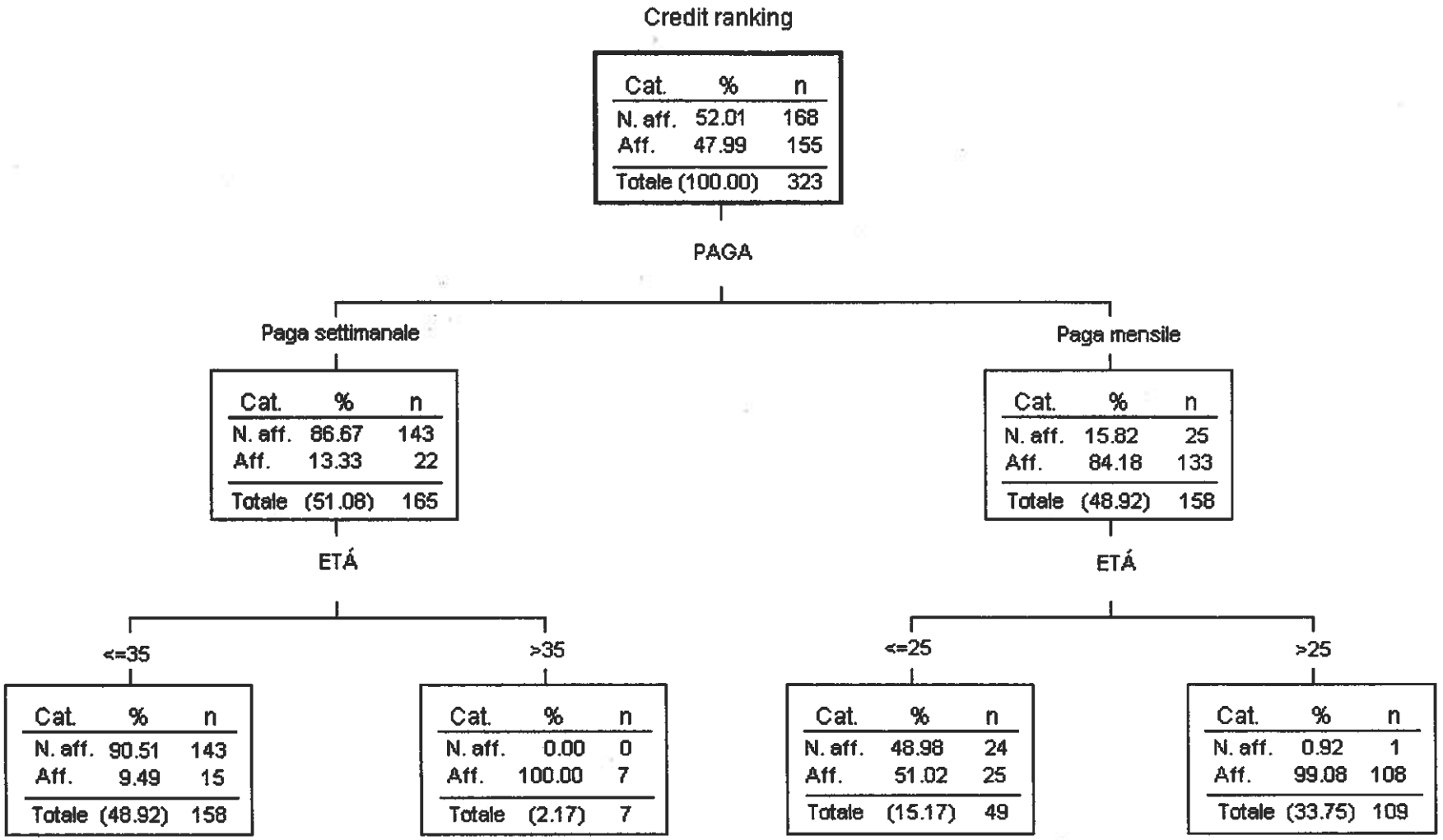
I clienti d'un istituto bancario americano sono stati classificati in base alla loro affidabilità nel pagamento delle quote d'un prestito elargito (variabile CREDIT RANKING). La modalità «affidabile» è stata attribuita a coloro che hanno rispettato tutte le scadenze del prestito, mentre nella categoria «non affidabile» sono stati inseriti i clienti che non hanno rispettato le scadenze e sono perciò caduti in mora. La classificazione è stata effettuata considerando la periodicità (mensile o settimanale) attraverso la quale viene percepita la retribuzione (PAGA) e l'età del cliente (ETÀ).

---

(2) Con l'espressione anglosassone *credit scoring* si intende definire un insieme di tecniche utilizzate dagli istituti di credito per la valutazione dell'affidabilità dei clienti basate sull'utilizzo di strumenti statistici.

(3) I dati sono stati estratti dal *file* credit.sav collocato nella *directory* di lavoro del modulo AnswerTree del *package* SPSS.

FIG. 7.1. *Classificazione di 323 clienti d'un istituto di credito secondo la rappresentazione ad albero.*



I dati suddetti possono essere presentati in una tabella a tripla entrata, la cui lettura non è però immediata (tab. 7.1). Per tale motivo nelle metodologie di segmentazione si ricorre ad una rappresentazione grafica detta «ad albero». Nella fig. 7.1 viene rappresentato il contenuto della tab. 7.1 secondo una struttura gerarchica.

Dal punto di vista formale, un albero rappresenta un insieme finito di elementi detti *nodi*. Il nodo da cui si diramano i successivi viene detto *radice* e verrà indicato nel seguito con la lettera R. L'insieme dei nodi, ad eccezione del nodo radice, può essere suddiviso in h insiemi distinti $S_1, S_2, ..., S_h$ che vengono indicati come *sottoalberi* del nodo R. L'insieme dei nodi discendenti da un determinato nodo intermedio viene denominato *branca*. Un nodo viene chiamato *padre* rispetto ai nodi che esso genera, mentre viene denominato *figlio* rispetto al nodo da cui discende. I valori di soglia d'una variabile che dividono le unità d'un determinato nodo sono chiamati *split*. I nodi terminali sono denominati *foglie* (4). L'insieme di tutti i nodi terminali d'un albero viene indicato con il simbolo $\tilde{T}$.

Nella fig. 7.1 il rettangolo superiore rappresenta la radice. Al suo interno sono riportate le modalità della variabile CREDIT RANKING con l'indicazione delle corrispondenti frequenze, che coincidono con i totali di colonna della tab. 7.1. Il valore percentuale dei totali riportato fra parentesi fornisce il peso d'un nodo nel livello a cui appartiene, che nel caso del nodo radice è ovviamente uguale al 100%. I due sottoalberi sottostanti sono formati distinguendo le due modalità secondo le quali si manifesta la variabile PAGA. A sinistra vengono collocati i clienti che percepiscono una retribuzione con cadenza settimanale, a destra abbiamo invece i clienti che vengono retribuiti mensilmente. Nell'esperienza statunitense (alla quale si riferisce l'esempio) la retribuzione settimanale è sintomo d'un lavoro instabile che può ripercuotersi negativamente sulla regolarità dei pagamenti alla banca. Infatti, nel nodo di sinistra compare una percentuale elevata (86.67%) di clienti non affidabili, mentre nel nodo di destra la maggioranza dei clienti (84.18%) risulta affidabile. Nelle foglie compaiono le unità statistiche classificate ulteriormente in base alla variabile ETÀ utilizzando come

(4) Gli alberi rientrano nella classe dei *grafi*. Nella terminologia specifica della teoria dei grafi i nodi vengono chiamati *vertici*, mentre le linee che uniscono i singoli nodi vengono definiti *archi* (per approfondimenti sul concetto di grafo si veda, ad esempio: Tutte, 1984).

split il valore 35 a sinistra ed il valore 25 a destra. Le frequenze riportate all'interno di ogni foglia corrispondono alle singole celle della tab. 7.1. La distinzione fra clienti affidabili e non affidabili ottenuta al livello finale è più netta rispetto alla classificazione ottenuta al primo livello. Fra tutti i clienti che percepiscono una retribuzione settimanale coloro che hanno un'età superiore ai 35 anni sono affidabili, mentre il 90.51% dei clienti con età inferiore o pari a 35 anni è inaffidabile. Si noti comunque che il peso della foglia contenente i 7 clienti affidabili aventi paga settimanale è molto basso perché rappresenta solo il 2.15% dei clienti complessivi. L'ulteriore ripartizione dei clienti con retribuzione mensile individua i clienti con età superiore ai 25 anni che sono quasi totalmente affidabili (99.08%). Se l'età è invece inferiore ai 25 anni (sempre per clienti con paga mensile) si verifica un equilibrio quasi perfetto fra affidabili e non affidabili. Una distinzione più netta potrebbe essere ottenuta utilizzando un'ulteriore variabile esplicativa.

Le conclusioni che si possono trarre da una classificazione gerarchica hanno conseguenze rilevanti in ambito previsivo. In base all'esperienza rappresentata dall'analisi dei 323 clienti, in futuro l'istituto di credito sarà ben disposto nella concessione d'un credito ad un potenziale cliente con retribuzione mensile e con età superiore ai 25 anni, mentre sarà costretto a richiedere una serie di garanzie a clienti giovani con retribuzione settimanale.

Lo scopo del presente capitolo è quello di illustrare le diverse tecniche che sono state proposte in letteratura per la creazione dei segmenti finali (5). Il contributo statistico più rilevante è sicuramente costituito dalla metodologia CART (*Classification And Regression Trees*; Breiman *et al.*, 1984). In tale lavoro viene introdotta la distinzione fra alberi di classificazione, in cui la variabile dipendente è di tipo categorico, e alberi di regressione, nei quali la variabile dipendente è di tipo

---

(5) Tali tecniche traggono origine dai lavori di Belson (1959) e di Morgan e Sonquist (1963). In particolare, nell'articolo di Morgan e Sonquist (1963) viene introdotta una procedura sequenziale detta AID (*Automatic Interaction Detection*) per l'individuazione automatica delle interazioni fra le variabili e per la classificazione delle unità statistiche indipendentemente dalle assunzioni di linearità delle relazioni. Successivamente, è stata proposta una variazione del metodo AID originale attraverso l'utilizzo del test Chi-quadrato (CHAID: *Chi-square Automatic Interaction Detection*; Kass, 1980).

Per un'esposizione in italiano delle tecniche di segmentazione si veda anche Fabbris (1997, pp. 355-396).

quantitativo. Più recentemente, sono state proposte alcune variazioni del metodo CART che sviluppano alberi non binari (Loh e Vanichsetakul, 1988; Keptra, 1996) o che riducono i tempi di calcolo (Mola e Siciliano, 1997). Infine, Loh e Shih (1997) hanno introdotto una nuova procedura (QUEST: *Quick, Unbiased, Efficient, Statistical Tree*) che trae spunto dalla metodologia CART, ma ne evita le distorsioni nella fase di selezione delle variabili.

In questo capitolo, dopo una breve introduzione sulla simbologia utilizzata, verranno presentate le più note tecniche di segmentazione con particolare riferimento alle fasi di costruzione d'un albero ed alle possibilità applicative legate al *package* statistico SPSS (modulo *AnswerTree*).

2. *Definizioni e notazioni*

Si consideri una variabile dipendente Y che presenta J modalità se qualitativa ovvero è suddivisa in J classi se quantitativa. Si considerino inoltre p variabili esplicative, quantitative o qualitative, $X_1, X_2, \ldots, X_p$ rilevate su n unità statistiche. Si indichi con $\mathbf{x}_i = [x_{i1}, \ldots, x_{is}, \ldots, x_{ip}]'$ il vettore contenente le informazioni per l'i-esima unità statistica (valori numerici per le variabili quantitative, codici per le variabili qualitative). La segmentazione può essere definita come una procedura «per passi» (*stepwise*) attraverso la quale l'insieme delle n unità viene suddiviso progressivamente, secondo un criterio di ottimizzazione, in una serie di sottogruppi disgiunti e che presentano al loro interno un grado di omogeneità maggiore rispetto all'insieme iniziale. La segmentazione fornisce pertanto una successione gerarchica di partizioni dell'insieme delle n unità ottenuta con un criterio scissorio o *top down* (v. cap. 5). Ad ogni passo del processo l'eterogeneità nei gruppi si riduce rispetto al passo precedente. Al termine, le foglie dell'albero, utilizzato per descrivere graficamente il procedimento, presentano un grado di omogeneità tale da poterle attribuire ad una delle J classi di partenza. Come si può intuire, tale tecnica, se n è grande, richiede una notevole mole di calcoli, che non potrebbe essere effettuata senza l'ausilio dei calcolatori elettronici (6).

---

(6) Infatti «*The tree methodology [...] is a child of the computer age. Unlike many other statistical procedures which were moved from pencil and paper to calcula-*

La segmentazione viene effettuata sulle n osservazioni. Essa conduce però ad individuare una regola che consente di classificare nuove osservazioni in una delle J classi della variabile Y. La costruzione d'un albero mediante una procedura di segmentazione definisce un criterio mediante il quale si assegna un'unità statistica ad una delle J classi della variabile Y. Tale regola potrà poi essere utilizzata per classificare nuovi casi di cui non si conosce la classe di appartenenza.

Si indichi con $\mathbf{X} \subseteq R^p$ lo spazio dei valori che possono assumere le p variabili (7).

Siano A_j $(j = 1, ..., J)$ le classi d'una partizione.

Definizione. Una *regola di classificazione* è una partizione di $\mathbf{X}$ in J sottoinsiemi $A_1, A_2, ..., A_J$, tale che per ogni $\mathbf{x} \in A_j$ la classe prevista è j, cioè

$$A_j = \{\mathbf{x}; d(\mathbf{x}) = j\}, \tag{7.1}$$

Nell'esempio riportato nel precedente paragrafo, $J = 2$, corrispondente al numero di modalità (affidabile, non affidabile) della variabile CREDIT RANKING. Per stabilire la classe alla quale attribuire le unità statistiche, utilizziamo la regola secondo la quale assegnamo alle foglie finali la classe corrispondente alla modalità più frequente. Secondo tale criterio, nella fig. 7.1 le unità statistiche della prima foglia a sinistra vengono assegnate alla classe «Non affidabile», quelle appartenenti alla seconda foglia vengono attribuite alla classe «Affidabile» e così via. La regola di decisione avrà, quindi la seguente forma:

$$A_1 = \{\mathbf{x}; d(PAGA\ sett.\ \text{e}\ ET\acute{A} \leq 35) = Non\ affidabile\}$$

$$A_2 = \{\mathbf{x}; d[(PAGA\ sett.\ \text{e}\ ET\acute{A} > 35)\ \text{o}\ (PAGA\ mens.)] = Affidabile\}.$$

Pertanto, a titolo d'esempio, un nuovo cliente della banca che percepisce una paga settimanale ed ha un'età di 28 anni sarà classificato «Non affidabile», mentre un cliente che percepisce una paga mensile (di qualunque età) sarà classificato «Affidabile».

---

tors and then to computers, this use of trees was unthinkable before computers. » (Breiman *et al.*, 1984).

(7) $\mathbf{X}$ è un sottospazio di R^p poiché alcune variabili possono assumere valori in un intervallo limitato, ovvero possono presentare solo un numero finito di modalità.

3. *Le fasi d'una procedura di segmentazione*

La definizione d'una procedura di segmentazione richiede l'impiego d'un insieme di strumenti decisionali che possono essere sintetizzati come segue:

1) dicotomizzazione delle variabili esplicative;
2) scelta del criterio di suddivisione di ogni nodo t nei nodi figli t_1 e t_2;
3) definizione d'un criterio di arresto nella costruzione dell'albero;
4) individuazione d'una regola per l'assegnazione d'una delle J modalità della variabile dipendente ad ogni foglia;
5) costruzione della regola $d(\mathbf{x})$ per la classificazione dei nuovi casi;
6) stima del tasso di errata classificazione.

3.1. *Dicotomizzazione delle variabili esplicative*

I criteri di partizione hanno come obiettivo quello di individuare la migliore suddivisione dello spazio $\mathbf{X}$ delle variabili esplicative ai fini della previsione della classe j della variabile dipendente. Per raggiungere tale obiettivo è necessario identificare tutte le possibili partizioni fra cui scegliere quella ottimale. Le possibili suddivisioni dipendono dalla natura quantitativa o qualitativa dei predittori (per una rassegna esaustiva sull'argomento, si veda Hawkins e Kass, 1982). Ai fini dell'analisi possiamo distinguere le variabili esplicative in (Vol. I, pp. 17-23):

— quantitative (ad es.: numero di dipendenti, fatturato);
— ordinali (ad es.: giudizio sulla solvibilità dell'azienda formulato da un gruppo di esperti);
— nominali (ad es.: forma giuridica delle aziende);
— dicotomiche (ad es.: principale mercato di sbocco delle aziende con modalità nazionale ed estero).

Se X_s è una variabile continua, la divisione binaria delle unità statistiche avviene individuando un valore che rappresenta la soglia di ripartizione; con riferimento a tale soglia si considerano due sottoinsiemi, da un lato quello comprendente tutti i valori inferiori o uguali al valore considerato e dall'altro tutti i valori superiori a quello considerato.

In generale, se nel campione la variabile X_s assume n valori distinti, la soglia corrisponde ad uno di questi valori (ad esclusione dell'ultimo) nella serie dei valori ordinati in senso non decrescente; in tale modo i possibili *split* connessi a quella variabile saranno pari a $n - 1$.

Nel caso d'una variabile ordinale ad m modalità, il numero di suddivisioni possibili sarà pari a $m - 1$. Se, ad esempio, distinguiamo i giudizi riportati da n individui in una prova d'esame secondo le modalità insufficiente, sufficiente, buono e ottimo, il numero di possibili *split* sarà pari a 3:

(insufficiente), (sufficiente, buono, ottimo)
(insufficiente, sufficiente), (buono, ottimo)
(insufficiente, sufficiente, buono), (ottimo).

Il caso delle variabili nominali ad m modalità è il più complesso perché non è possibile stabilire un ordinamento. Utilizzando le nozioni del calcolo combinatorio, si determina il numero delle possibili suddivisioni, che è pari a $2^{m-1} - 1$. Quindi, se consideriamo la variabile «forma giuridica» di n aziende con modalità «azienda individuale», «società di persone», «società di capitale», «altro», il numero di possibili partizioni con due classi sarà pari a $2^{4-1} - 1 = 7$, cioè:

(az. individuale, soc. di persone), (soc. di capitale, altro)
(az. individuale, soc.di capitale), (soc. di persone, altro)
(soc. di persone, soc. di capitale), (az. individuale, altro)
(az. individuale, soc. di persone, soc. di capitale), (altro)
(az. individuale, soc. di persone, altro), (soc. di capitale)
(az. individuale, soc. di capitale, altro), (soc. di persone)
(soc. di persone, soc. di capitale, altro), (az. individuale).

Si noti che al crescere di m il numero di possibili *split* aumenta in modo più che proporzionale e può diventare molto elevato.

La variabile dicotomica è chiaramente un caso particolare di variabile nominale in cui $m = 2$; la suddivisione corrispondente è unica poiché i sottogruppi che si possono individuare corrispondono alle due modalità assumibili dalla variabile (ad esempio, Affidabile e Non affidabile per la variabile CREDIT RANKING) (8).

3.2. *Criterio di suddivisione d'un nodo*

La fase centrale d'una procedura di segmentazione è senza dubbio la suddivisione delle unità appartenenti ad un nodo e di conseguenza la scelta del criterio in base al quale effettuare tale ripartizione. Come

(8) Oltre a suddivisioni basate su singole variabili è possibile utilizzare *split* definiti su combinazioni lineari di variabili continue o su combinazioni di modalità di variabili qualitative.

verrà esposto nei prossimi paragrafi, è proprio questa fase che distingue tra loro le tecniche di classificazione gerarchica proposte in letteratura.

Un criterio di suddivisione (o di *split*) consiste nel calcolo d'un indice statistico che consenta di selezionare la partizione migliore fra tutte le possibili, corrispondenti ad ogni singolo predittore. Fra tutti i predittori verrà poi selezionato il migliore in relazione al criterio di riduzione dell'eterogeneità prescelto. La bontà di tale criterio deve poi essere valutata attraverso il concetto di coerenza, per cui l'insieme iniziale deve essere suddiviso in gruppi il più possibile omogenei al loro interno e il più possibile eterogenei fra loro (9).

Generalmente, gli algoritmi di segmentazione consistono nella ricerca del migliore *split* analizzando tutte le p variabili esplicative.

3.3. *Criterio di arresto*

Le tecniche di segmentazione gerarchica, come s'è detto, consistono nella ripartizione ricorsiva d'un insieme di unità statistiche. Tale metodologia ricorsiva richiede la definizione d'una o più regole di stop al verificarsi delle quali il processo si blocca. In caso contrario la segmentazione si arresta quando i nodi terminali contengono solo casi appartenenti alla medesima classe della variabile dipendente.

Le proprietà desiderabili d'una regola di arresto sono la semplicità e il potere discriminatorio. In base alla prima proprietà, fra due regole di arresto si sceglie, *ceteris paribus*, quella che determina l'albero di taglia minore e quindi più facilmente leggibile in fase di interpretazione dei risultati. La seconda proprietà riguarda invece l'esigenza di ottenere strutture decisionali che permettano di distinguere nel modo più efficace possibile unità statistiche appartenenti a classi diverse. Come è facilmente intuibile, le due proprietà sono tra loro opposte e difficilmente conciliabili.

Le tecniche di segmentazione più note utilizzano solitamente regole di arresto basate sulla numerosità minima dei nodi terminali, o su livelli massimi consentiti per la crescita dell'albero. Il metodo CART proposto da Breiman *et al.* (1984) utilizza invece una strategia originale alternativa basata sul concetto di potatura (in inglese *pruning*). Essa si articola in due fasi: dapprima si costruisce l'albero di massima dimen-

---

(9) Si noti l'analogia con i criteri di coesione interna e separazione esterna per la formazione dei gruppi nella *cluster analysis*, illustrati nel cap. 5.

sione che contiene in ogni nodo un solo elemento oppure elementi appartenenti alla stessa classe e successivamente si sfronda l'albero massimo secondo una regola che minimizza la complessità a parità di potere discriminatorio.

3.4. *Assegnazione delle classi alle foglie e classificazione di nuovi casi*

Quando il processo di costruzione è terminato è necessario stabilire quale classe corrisponde ad ogni nodo terminale. A tale fine è necessario distinguere almeno tre situazioni:

— la foglia comprende casi appartenenti ad una sola classe; chiaramente la classe assegnata alla foglia è quella corrispondente alle unità che ne fanno parte (regola dell'unanimità);

— nella foglia sono presenti unità statistiche di classe diversa, ma una di queste ha frequenza superiore alle altre; la classe della foglia corrisponde a quella con frequenza massima; tale regola viene definita regola della maggioranza o *plurality rule* (Breiman *et al.*, 1984, p. 26);

— le unità della foglia appartengono a classi diverse con medesima frequenza; in questo caso si cade in una zona di indecisione. Tali situazioni di indecisione sono rare quando vengono applicate le tecniche di *pruning*.

Dopo avere assegnato una classe ad ogni singolo nodo terminale è possibile procedere alla classificazione di nuovi casi al di fuori del campione utilizzato per la costruzione dell'albero. Applicando la regola di classificazione dell'albero, ogni singolo caso ricade in una foglia e viene etichettato in base alla classe assegnata alla foglia corrispondente.

3.5. *La stima del tasso di errata classificazione*

La scelta della migliore regola di classificazione avviene attraverso una misura della bontà di assegnazione delle unità statistiche. A parità di semplicità della rappresentazione ad albero (in termini di numero di foglie) verrà selezionata la regola che consente di allocare correttamente la percentuale più elevata di unità statistiche. La misura utilizzata per valutare la bontà del classificatore è il *tasso di errata classificazione* associato alla regola d indicato con il simbolo $R(d)$.

Definiamo con S il campione di unità statistiche in relazione al quale viene costruita la regola di classificazione d e con Ω un insieme artificiale molto numeroso (virtualmente infinito) di unità statistiche

avente le stesse caratteristiche di S. Teoricamente, il tasso di errata classificazione dovrebbe essere calcolato confrontando la reale classificazione delle osservazioni in Ω con quella prevista da d. Tale procedura non può essere applicata nei casi concreti per cui è necessario ricorrere ad una stima di $R(d)$ che chiameremo $\widehat{R}(d)$. In letteratura sono stati proposti diversi metodi di stima di $R(d)$ che verranno di seguito esposti.

Stima basata sul campione S (resubstitution estimate). Si definisca con $C_j(i)$ la classe di effettiva appartenenza della i-esima unità statistica e con $d(\mathbf{x}_i)$ la classe assegnata alla stessa unità statistica dalla regola d. Sia inoltre $I(\cdot)$ una funzione indicatrice che assume valore 1 se l'affermazione all'interno delle parentesi è vera e valore 0 nel caso contrario. La stima per risostituzione (basata sul campione) del tasso di errata classificazione corrispondente al classificatore d è quindi:

$$\widehat{R}(d) = \frac{1}{n}\sum_{i=1}^{n} I[d(\mathbf{x}_i) \neq C_j(i)] \tag{7.2}$$

È stato constatato empiricamente che $\widehat{R}(d)$, pur rappresentando un metodo computazionalmente semplice, fornisce stime ottimistiche di $R(d)$. Infatti, si effettua un test sulla regola di classificazione utilizzando gli stessi dati sui quali è stata costruita la regola stessa.

Stima basata su un campione test (test sample estimate). Il campione S viene partizionato in maniera casuale in due sottocampioni S_1 e S_2 di numerosità, rispettivamente, n_1 e n_2, tali che $S_1 \cup S_2 = S$ e $S_1 \cap S_2 = \phi$ (dove ϕ indica l'insieme vuoto). S_1 viene definito campione di apprendimento (*learning sample*), S_2 viene denominato campione test (*testing sample*). Per la tecnica di suddivisione adottata, i due campioni possono considerarsi indipendenti. La regola d viene costruita utilizzando il campione d'apprendimento e viene testata stimando $R(d)$ sul campione test, per cui la stima del campione test, $\widehat{R}_{ts}(d)$, è data da

$$\widehat{R}_{ts}(d) = \frac{1}{n_2}\sum_{i \in S_2} I[d(\mathbf{x}_i) \neq C_j(i)] \tag{7.3}$$

Le stime del tasso di errata classificazione ottenute applicando questo metodo sono più affidabili rispetto alle precedenti perché utilizzano dati esterni rispetto a quelli impiegati per la costruzione della regola di classificazione. I limiti della procedura risiedono nella riduzione del

campione su cui viene costruito il classificatore d, per cui è auspicabile la sua applicazione solo se il campione S è di numerosità elevata.

Stima basata sulla cross-validation (V-fold cross-validation estimate). Tale metodo di stima è preferibile alla stima basata sul campione test, qualora la numerosità di S sia bassa. Il campione iniziale S viene ripartito casualmente in $V > 2$ sottocampioni $S_1, S_2, ..., S_v, ..., S_V$ di dimensione il più possibile prossima fra di loro. La regola di classificazione $d^{(v)}(\mathbf{x})$ viene costruita sul campione $S - S_v$, $v = 1, 2, ..., V$. Poiché per ogni v nessun elemento di S_v è compreso in $S - S_v$, è possibile effettuare una stima basata sul campione test di $R(d^{(v)})$, cioè

$$\hat{R}_{ts}(d^{(v)}) = \frac{1}{n_v} \sum_{i \in S_v} I\left[d^{(v)}(\mathbf{x}_i) \neq C_j(i)\right], \qquad v = 1, 2, ..., V. \tag{7.4}$$

Se V è sufficientemente elevato, ogni classificatore $d^{(v)}(\mathbf{x})$ viene costruito utilizzando un campione di apprendimento di dimensione $n(1 - 1/V)$ che è prossima alla dimensione di S. L'assunzione portante del metodo di *cross-validation* è quindi la stabilità, nel senso che ogni classificatore $d^{(v)}(\mathbf{x})$, $v = 1, 2, ..., V$ ha un tasso di errata classificazione $R(d^{(v)})$ molto prossimo a $R(d)$. La stima del tasso di errata classificazione basata sulla *cross-validation* $\hat{R}_{cv}(d)$ ha quindi la seguente forma:

$$\hat{R}_{cv}(d) = \frac{1}{V} \sum_{v=1}^{V} \hat{R}_{ts}(d^{(v)}). \tag{7.5}$$

4. *Metodi di segmentazione con variabili esplicative qualitative: AID e CHAID*

La tecnica di segmentazione più famosa in letteratura è quella nota come AID (*Automatic Interaction Detection*) proposta da Morgan e Sonquist (1963). Tale metodologia fu introdotta per risolvere problemi di analisi di campioni di grande dimensione rispetto ai quali viene rilevato un numero elevato di variabili. La tecnica AID può essere applicata quando la variabile dipendente è quantitativa e le variabili esplicative sono caratteri qualitativi. Il criterio di suddivisione dei nodi è di tipo binario e si basa essenzialmente sulla scomposizione della varianza della variabile dipendente nella quota entro e fra i gruppi.

Kass (1980) ha ripreso molti concetti esposti nell'ambito della procedura AID, proponendo un metodo di segmentazione alternativo ba-

sato sul test del chi-quadrato denominato CHAID (*Chi-square Automatic Interaction Detection*).

CHAID si differenzia dalla AID per i seguenti motivi:

— la variabile dipendente è qualitativa anziché quantitativa;

— la scelta del predittore in base al quale eseguire lo *split* ad un determinato livello dell'albero si basa su un test statistico, anziché sulla scomposizione della varianza; ciò permette di tenere in considerazione la variabilità campionaria che veniva trascurata dal metodo AID (Bishop *et al.*, 1975, p. 360);

— sono consentiti gli *split* multipli, cioè un nodo genitore può generare più di due nodi figli; tale caratteristica rende la tecnica CHAID preferibile rispetto ad altri metodi di segmentazione qualora si voglia superare il limite della suddivisione binaria;

— viene definita una nuova categoria di variabili esplicative (denominate variabili *floating*) ed il metodo da utilizzare per il loro trattamento; tale categoria è costituita da variabili qualitative per le quali è possibile stabilire un ordinamento fra tutte le modalità ad eccezione di una; un caso notevole di variabili *floating* è quello di una variabile espressa su scala ordinale, le cui modalità non sono disponibili per alcune unità statistiche (dati mancanti).

4.1. *Accorpamento delle modalità delle variabili esplicative*

Come tutti i metodi di segmentazione, CHAID procede secondo stadi successivi. Dapprima si seleziona la migliore suddivisione delle unità statistiche per ogni predittore. Successivamente, tutti i predittori vengono confrontati fra loro al fine di scegliere il migliore in termini di omogeneità della suddivisione determinata; le unità appartenenti al nodo vengono ripartite in base alla variabile esplicativa selezionata. Infine, ogni sottogruppo viene ripreso in considerazione in modo indipendente dagli altri sottogruppi per individuare eventuali ulteriori suddivisioni.

Data una variabile dipendente Y che si manifesta secondo $J \geq 2$ modalità e un particolare predittore che può assumere $C \geq 2$ modalità, è possibile rappresentare la loro distribuzione di frequenze congiunta mediante una tabella di contingenza $C \times J$. L'obiettivo centrale della procedura CHAID è quello di ridurre la tabella di contingenza iniziale ad una tabella di contingenza $C' \times J$, $(C' < C)$ accorpando alcune modalità della variabile esplicativa. Il criterio utilizzato per l'accorpamento si basa sul concetto di indipendenza in una tabella di contingenza (per

un approfondimento di tale argomento, rinviamo al vol. I, cap. VI): fra tutti i possibili modi di accorpare le modalità della variabile esplicativa si seleziona quello che determina il più elevato grado di associazione con le classi della variabile dipendente, purché superi una determinata soglia. In tale modo si uniscono le modalità del predittore che statisticamente sono tra loro più simili.

Si indichi con $\chi^2_{C'}(u)$ il valore campionario della statistica χ^2 (Vol. I, appendice al capitolo VI) per l'u-esimo modo di accorpamento delle modalità in una tabella $C' \times J$ ($C' = 2, 3, ..., C$). Il campo di variazione di u dipende dalla natura del predittore (nominale, ordinale, *floating*). Infatti, il numero di modi in cui le C modalità di una variabile qualitativa possono essere raggruppate in C' classi è diverso per una variabile ordinale, in cui solo categorie contigue possono essere raggruppate fra di loro, rispetto, per esempio, ad una variabile nominale. Le formule per il calcolo esatto del *range* di u vengono riportate nell'articolo di Kass (1980).

4.2. *Il criterio di* split

Il valore obiettivo di $\chi^2_{C'}(u)$ viene determinato attraverso una procedura di tipo *stepwise* i cui stadi sono i seguenti.

1) Per ogni predittore si costruisce la tabella a doppia entrata rispetto alla variabile dipendente. Si eseguono, quindi, gli *step* 2 e 3.

2) Per ogni possibile coppia di modalità della variabile dipendente si calcola il valore della statistica χ^2 per verificare l'ipotesi nulla di indipendenza tra le coppie di modalità e la variabile dipendente. Si seleziona la coppia cui corrisponde il più basso valore di χ^2. Se, per la coppia selezionata, l'ipotesi nulla non può essere rifiutata ad un livello di significatività α_m, le due modalità considerate vengono fuse in un'unica categoria e si passa allo stadio 3. Nel caso in cui il χ^2 assuma un valore superiore alla soglia corrispondente a α_m si passa allo stadio 5.

3) Si verifica in corrispondenza di ogni classe composta da tre o più modalità originarie del predittore, se essa è disaggregabile utilizzando lo stesso criterio del χ^2. Se l'ipotesi di indipendenza può essere rifiutata ad un livello di significatività α_s si effettua lo *split* e si ritorna al passo 2. In pratica un'unione verrà raramente divisa, ma tale possibilità deve comunque essere presente per il raggiungimento d'una soluzione quasi-ottimale.

4) Si calcola per ogni predittore la significatività della ripartizione (attraverso il p-value) ottenuta ai passi precedenti. Nel caso in esame, il

p-value è la probabilità che, sotto l'ipotesi nulla di indipendenza tra variabile dipendente e predittore, si osservi un valore di χ^2 superiore a quello ottenuto. Se non sono state effettuate aggregazioni sulla tabella di contingenza originaria, la significatività può essere calcolata attraverso il consueto valore del χ^2. Nel caso in cui siano state effettuate delle aggregazioni di modalità, la significatività della ripartizione deve essere calcolata considerando congiuntamente tutte le combinazioni di modalità che sono state testate per il predittore considerato (10).

5) Per ogni partizione dei dati non ancora analizzata si torna al passo 1.

4.3. *Il criterio d'arresto*

I passi esposti nel paragrafo precedente vengono eseguiti iterativamente, perciò giunti all'ultimo stadio si torna al primo per analizzare ogni nodo figlio che contiene un numero di osservazioni superiore o uguale alla dimensione minima prefissata. L'algoritmo si arresta quando tutti i nodi terminali contengono elementi della stessa classe o un numero di casi inferiore alla soglia (11).

---

(10) Poiché le sottotabelle di dimensione $C' \times J$ che sono state testate dall'algoritmo sono fra loro dipendenti non è possibile utilizzare il metodo classico per il calcolo del *p-value* congiunto per test multipli in caso di indipendenza.

Date n tabelle di contingenza tra loro indipendenti in relazione alle quali viene effettuato un test χ^2 di indipendenza con una probabilità α di commettere errore di prima specie, il *p-value* congiunto α' sarà:

$$\alpha' = 1 - (1 - \alpha)^n.$$

Bonferroni (1936) ha proposto un metodo approssimato per il calcolo del limite inferiore del *p-value* per test multipli dipendenti. Nel caso dei test multipli di indipendenza per tabelle di contingenza, tale limite inferiore si ottiene moltiplicando il *p-value* ottenuto in caso di indipendenza per B, dove B è il fattore di correzione di Bonferroni ed è pari al numero di modi in cui si possono combinare le modalità del predittore. Il predittore con il valore minimo del *p-value* corretto, purché inferiore al valore soglia, viene selezionato e le osservazioni vengono ripartite secondo le categorie aggregate di tale predittore.

(11) È stato proposto in letteratura (Biggs *et al.*, 1991) un avanzamento di CHAID chiamato *Exaustive CHAID*. Il maggiore contributo metodologico di tale procedura consiste nella correzione della distorsione che CHAID presenta a favore di partizioni semplici ad ogni passo della segmentazione.

4.4. *Un esempio*

A titolo d'esempio, prendiamo in considerazione un insieme di dati tratti da SPSS (*file*: impiegati.sav) riguardante gli stipendi di 474 dipendenti d'una grande azienda in relazione ad alcune variabili quali la retribuzione iniziale, il livello di istruzione, il sesso. Per applicare la metodologia CHAID è stato necessario trasformare le variabili quantitative in variabili qualitative, dividendo il loro campo di variazione in intervalli, poi codificati. È stata selezionata come variabile dipendente la retribuzione (in migliaia di dollari statunitensi) che i dipendenti percepivano al momento dell'indagine (STIPATT) suddivisa nelle seguenti classi: ≤ 25 (codice 1), da più di 25 a 35 (codice 2), >35 (codice 3). Le restanti variabili sono state utilizzate come predittori nella costruzione dell'albero:

— data di nascita (NASCITA) con le seguenti modalità: prima del 1945 (codice 1), fra il 1945 e il 1960 (codice 2), dopo il 1960 (codice 3);

— livello di istruzione (STUDIO) con le seguenti modalità: scuola dell'obbligo (codice O), scuola superiore (codice S), laurea (codice L), scuola di specializzazione post-laurea o titolo equivalente (codice LL);

— stipendio iniziale (in migliaia di dollari) (STIPINIZ) con le seguenti modalità: ≤ 15 (codice 1), da più di 15 a 20 (codice 2), >20 (codice 3);

— anni trascorsi dalla data di assunzione in azienda (ANNILAV) con le seguenti modalità: 5, 6, 7, 8;

— durata delle esperienze di lavoro precedenti (in mesi) (ESPPREC) con le seguenti modalità: ≤ 36 (codice 1), da più di 36 a 120 (codice 2), >120 (codice 3);

— categoria lavorativa di appartenenza (CATLAV) con le seguenti modalità: impiegato (codice 1), funzionario (codice 2), dirigente (codice 3);

— sesso (SESSO) con le modalità maschio (codice *m*) e femmina (codice *f*).

I passaggi necessari per la costruzione dell'albero attraverso il modulo *AnswerTree* di SPSS sono riportati di seguito.

1) Dalla maschera principale si sceglie la sequenza *file-new project* e si seleziona il nome del file che contiene i dati iniziali in formato leggibile da SPSS (nome file.sav). Il file selezionato deve contenere la variabile dipendente (qualitativa) e tutte le potenziali variabili esplicative (qualitative). Le modalità delle variabili devono essere opportunamente codificate.

2) Dal menù principale si seleziona la sequenza *file-new tree*. La variabile dipendente (nel caso in esame è la variabile STIPATT) deve essere collocata nella maschera *target*, mentre i potenziali predittori (tutte le variabili ad esclusione di STIPATT) devono essere disposti nella maschera *predictors*. Nello spazio riservato al *growing method* si seleziona CHAID, che è anche il metodo di *default*.

3) Premendo il tasto *OK* compare una nuova finestra in cui viene riportata la distribuzione di frequenza della variabile dipendente all'interno d'una cornice. Tale cornice rappresenta il nodo-radice dell'albero. Dal menù principale della nuova maschera si seleziona la sequenza *analysis-growing criteria*. È possibile definire due regole d'arresto comuni ad ogni metodologia di segmentazione: la profondità massima dell'albero (*maximum tree depth*) e il numero minimo di casi (*minimum number of cases*) presenti nel nodo genitore (*parent node*) e nel nodo figlio (*child node*). La prima regola d'arresto si riferisce al numero massimo di livelli gerarchici attraverso i quali viene costruita la segmentazione. Per evitare che tale criterio (piuttosto rozzo) assuma un peso eccessivo nella costruzione dell'albero è stata fissata una profondità massima pari a 10 livelli. Il numero minimo di casi per un nodo genitore significa che un nodo non può essere ulteriormente suddiviso se la sua numerosità è inferiore a tale limite (nell'esempio è stata posta pari a 50). Il numero minimo di casi per un nodo figlio impedisce invece la creazione di un nodo la cui numerosità sia ad esso inferiore (nell'esempio pari a 10). Specifica di CHAID è la scelta del livello di significatività α_m (*alpha for merging*) e del livello di significatività α_s (*alpha for splitting*) che nell'esempio sono stati fissati entrambi pari a 0.05. Aumentando α_m si riduce la probabilità di fondere fra di loro modalità diverse, mentre aumentando α_s si aumenta la probabilità di dividere modalità fuse ai passi precedenti.

4) Si suddivide il campione S nei due sottocampioni S_1 (*learning sample o training sample*) e S_2 (*testing sample*) selezionando la sequenza *analysis-partition data*. Nell'esempio 2/3 delle osservazioni (pari a 318) sono state incluse nel *learning sample* e le rimanenti (156) nel *testing sample*.

5) Selezionando la sequenza *tree-grow tree* il programma crea l'albero utilizzando i criteri di crescita che sono stati selezionati precedentemente.

Nella fig. 7.2 viene riportato l'albero costruito sul *learning sample*. Ad ogni *split* viene indicata la variabile in base alla quale è stata effettuata la suddivisione e il valore del test χ^2 corrispondente. Al di sopra

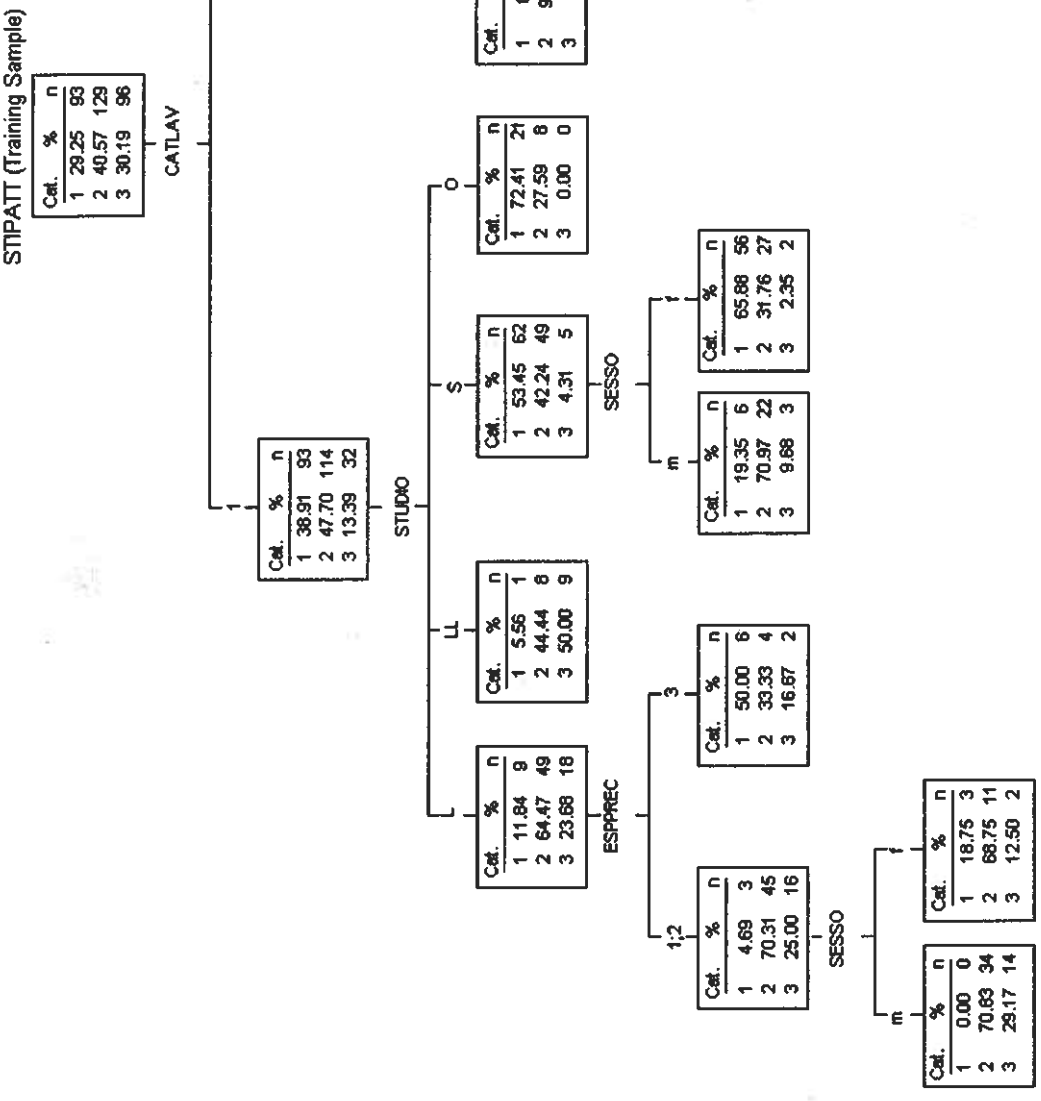


FIG. 7.2. *Albero della regola di classificazione per 474 impiegati ottenuto con la metodologia CHAID.*

di ogni nodo (ad esclusione della radice) sono riportate le modalità assunte dai casi che vi appartengono. Ad esempio, nel nodo di destra che si forma al secondo livello dell'albero, ricadono i funzionari e i dirigenti (modalità 2 e 3 della variabile CATLAV), mentre nel nodo di sinistra troviamo gli impiegati (modalità 1 della variabile CATLAV). È interessante notare che nel nodo di destra nessun dipendente ha una retribuzione inferiore a 25000 dollari. Al livello successivo la stessa variabile viene utilizzata per suddividere ulteriormente i dipendenti fra funzionari (modalità 2) e dirigenti (modalità 3): il 93.33% dei funzionari ha una retribuzione media (compresa fra 25000 e 35000 dollari), mentre il 98.44% dei dirigenti percepisce una retribuzione alta (maggiore di 35000 dollari). Gli impiegati presentano una retribuzione maggiormente variabile e la loro classificazione richiede l'utilizzo di ulteriori caratteri. L'interpretazione dei nodi è analoga alla precedente. In particolare, fra gli impiegati nessuno di coloro che sono in possesso d'una istruzione equivalente alla scuola dell'obbligo percepisce una retribuzione alta, mentre il 94.44% degli specializzati post-laurea ha una retribuzione medio-alta. Infine, si noti che alcune delle variabili introdotte nel modello (NASCITA, STIPINIZ, ANNILAV) non sono state selezionate per la segmentazione a causa del loro scarso potere discriminante. Classificando le unità del *testing sample* attraverso l'albero costruito sul *learning sample* si ottiene un tasso di errata classificazione $\widehat{R}_{ts} = 0.288$.

Il modello di segmentazione ottenuto nell'esempio è utile per comprendere la struttura di retribuzione dell'azienda, ma può essere utilizzato anche in chiave previsiva: un individuo che decide di avanzare una candidatura per l'assunzione in azienda può prevedere con una certa precisione il livello di retribuzione che ragionevolmente percepirà in base alle sue caratteristiche (sesso, titolo di studio, esperienze precedenti, categoria di assunzione).

5. *Alberi di classificazione e di regressione (CART)*

La metodologia CART (*Classification And Regression Trees*) proposta da Breiman *et al.* (1984) ha rappresentato un punto di svolta rispetto alle tecniche di segmentazione note in precedenza. Molti sono infatti gli elementi innovativi che possono essere sintetizzati come segue:

• la variabile dipendente può essere sia qualitativa, sia quantitativa; nel primo caso si ottiene un «albero di classificazione», nel secondo caso il modello viene denominato «albero di regressione»;

• è possibile considerare congiuntamente predittori qualitativi e quantitativi;

• gli *split* possono essere eseguiti considerando come predittori combinazioni lineari di variabili quantitative;

• il criterio di *split* viene definito in base al concetto di «impurità» d'un nodo; a differenza della metodologia CHAID non viene selezionata la variabile più significativa, ma quella che produce la massima riduzione di impurità;

• viene introdotto un metodo originale per il trattamento dei dati mancanti basato sul concetto di *surrogate split*;

• si propone il dimensionamento ottimale degli alberi di grossa dimensione attraverso una procedura di potatura (*pruning*).

A fronte degli elementi positivi elencati, la tecnica CART consente solo partizioni binarie. Tale limitazione è uno dei motivi per cui la metodologia CHAID può risultare preferibile a CART qualora le variabili esplicative siano tutte qualitative.

5.1. *Alberi di classificazione: criterio di* split *basato sul concetto di impurità*

Consideriamo dapprima gli alberi di classificazione. In tale caso la variabile dipendente è di tipo categorico con J modalità. L'idea di base per la creazione degli alberi di classificazione è di selezionare ogni suddivisione d'un insieme in modo tale che ciascuno dei sottogruppi prodotti dalla ripartizione sia più «puro» rispetto all'insieme di partenza. Il concetto di impurità si riferisce all'eterogeneità (Vol. I, pp. 117-138) delle unità statistiche in relazione alle modalità della variabile dipendente (12). In termini operativi, partendo dal nodo radice (o nodo padre) t si cerca la variabile che produce la migliore suddivisione degli n casi contenuti in t in due nodi figli (t_l e t_r) di numerosità n_l e n_r. I due nodi figli sono più omogenei rispetto al nodo padre.

---

(12) Dato un fenomeno qualitativo che può assumere r modalità, l'eterogeneità (impurità) è nulla se le n unità statistiche presentano la medesima modalità. Al contrario, l'eterogeneità è massima se le unità statistiche sono equamente ripartite fra le r modalità.

Si consideri l'albero riportato nella fig. 7.1. I due nodi intermedi che si ottengono suddividendo l'insieme in base alla modalità di retribuzione (mensile o settimanale) sono più puri rispetto al nodo padre perché all'interno di ciascuno di essi è diminuita l'eterogeneità della variabile dipendente. Infatti l'indice di Gini (13) calcolato sul nodo genitore risulta uguale a 0.499 e sui nodi figli di sinistra e di destra risulta pari, rispettivamente, a 0.231 e 0.266.

Per formalizzare il concetto, si indichi con $p(j|t)$ la proporzione dei casi di classe j presenti nel nodo t, con $j = 1,2,...,J$ e $p(1|t) + ... + p(J|t) = 1$.

Definizione. Si definisce *misura di impurità* associata ad un determinato nodo t la seguente funzione:

$$imp(t) = \phi[p(1|t), ..., p(j|t), ..., p(J|t)]$$

dove $\phi(\cdot)$ è una funzione non negativa tale che:

— $\phi[p(1|t), ..., p(j|t), ..., p(J|t)] = \max$, quando $p(j|t) = 1/J$ per $j = 1,2,...J$;

— $\phi[1,0,...,0,0] = 0$, $\phi[0,1,...,0,0] = 0, ...,$
$\phi[0,0,...,1,0] = 0$, $\phi[0,0,...,0,1] = 0$;

— è invariante rispetto all'ordine delle modalità.

Pertanto, l'impurità d'un nodo è massima quando tutte le classi della variabile dipendente sono presenti nella stessa proporzione, mentre è minima quando il nodo contiene casi appartenenti ad un'unica classe.

Diverse sono le funzioni di impurità utilizzate in letteratura. La più diffusa è l'indice di eterogeneità di Gini, cioè:

$$imp(t) = \sum_{j \neq j'} p(j|t)p(j'|t) = 1 - \sum_{j} p^2(j|t). \qquad (7.6)$$

---

(13) L'indice di eterogeneità di Gini è calcolato nel modo seguente:

$$G = 1 - \sum_{i=1}^{r} f_i^2$$

dove f_i è la frequenza relativa della modalità i-esima d'un fenomeno qualitativo che può assumere r modalità. G assume valore minimo (pari a 0) nel caso di massima omogeneità e valore massimo $(\frac{r-1}{r})$ nel caso di massima eterogeneità.

L'utilizzo preferenziale dell'indice (7.6) rispetto alle altre possibili misure di eterogeneità dipende, oltre che dalla sua relativa semplicità computazionale, dalla doppia interpretazione che ad esso è attribuibile. Infatti la prima uguaglianza della (7.6) rappresenta la stima della probabilità di errata classificazione di un'osservazione di classe j nella classe j', qualora l'assegnazione di un'unità del nodo t ad una particolare classe avvenga casualmente. La seconda uguaglianza è invece interpretabile in termini di varianza del nodo t qualora si codifichino con «1» i casi di classe j appartenenti al nodo t e con «0» i casi di classe diversa (14).

Definizione. Si definisce *misura del decremento di impurità* del nodo t associata ad un determinato *split* (s), la seguente quantità:

$$\Delta imp(s,t) = i(t) - p_l[imp(t_l)] - p_r[imp(t_r)], \tag{7.7}$$

dove p_l e p_r rappresentano la proporzione di casi del nodo t che cadono, rispettivamente, nel nodo di sinistra (*left*) e nel nodo di destra (*right*). La quantità $\Delta imp(s,t)$ è sempre non negativa e assume valore zero nella situazione estrema in cui $p(j|t_l) = p(j|t_r) = p(j|t)$, per $j = 1,2,...,J$.

Dopo aver creato tutte le possibili dicotomizzazioni delle variabili esplicative, coerentemente alla loro natura, gli alberi di classificazione vengono costruiti scegliendo, per un dato nodo t, lo *split* s^* che produce la massima riduzione di impurità dell'albero, cioè:

$$\Delta imp(s^*,t) = \max_{s\in\Theta} \Delta imp(s,t) \tag{7.8}$$

dove Θ è l'insieme di tutte le suddivisioni che si possono formare in relazione al nodo t. La scelta di s^* viene effettuata per ogni nodo e ad ogni livello dell'albero. Indicando con $IMP(t) = p(t)imp(t)$, l'impurità totale del generico albero T si definisce nel seguente modo:

---

(14) Clark e Pregibon (1992) e Ripley (1996, pp. 216-221) impostano il problema dell'impurità secondo un approccio diverso. L'albero viene infatti considerato come un modello probabilistico definito su un campione d'apprendimento. Quindi viene costruita una funzione di verosimiglianza basata sul modello probabilistico e si seleziona lo *split* che determina la massimizzazione d'una particolare misura di devianza del nodo.

$$IMP(T) = \sum_{t \in \tilde{T}} IMP(t) = \sum_{t \in \tilde{T}} imp(t)p(t) \qquad (7.9)$$

dove $p(t)$ rappresenta la proporzione di unità statistiche presenti nel nodo t e $\tilde{T}$ indica l'insieme dei nodi terminali. Si può dimostrare (Breiman *et al.*,1984, pp. 32-33) che la selezione dello *split* che massimizza il decremento di impurità $\Delta imp(s,t)$ è equivalente alla selezione dello *split* che minimizza l'impurità totale dell'albero. Ciò significa che il criterio di ottimizzazione locale d'un albero di classificazione equivale alla sua ottimizzazione globale (15).

---

(15) Ulteriori criteri di formazione d'un albero previsti da alcuni *packages* statistici (fra i quali *AnswerTree*) sono il *twoing* e l'*ordering twoing*. Tali metodi sono stati ideati ed introdotti in letteratura per il trattamento delle variabili dipendenti con un numero elevato di modalità. Si divide l'insieme A delle J modalità della variabile risposta in due sottoinsiemi:

$$A_1 = \{j_1, j_2, \ldots, j_m\} \qquad A_2 = A - A_1.$$

Per ogni possibile *split* s del nodo t si calcola $\Delta i(s,t)$ considerando l'appartenenza delle unità statistiche ad una delle due *superclassi* A_1 e A_2, cosicché il problema a più classi viene ridotto ad un problema con due sole classi. Poiché $\Delta i(s,t)$ dipende dalla suddivisione iniziale delle modalità nelle due *superclassi* A_1 e A_2, si utilizza la notazione $\Delta i(s,t,A_1)$. Quindi si determina il migliore *split* $s^*(A_1)$ in modo tale che

$$\Delta i[s^*(A_1), t, A_1] = \max_{s \in \Theta_{A_1}} \Delta i\,(s,t,A_1)$$

dove Θ_{A_1} è l'insieme di tutti gli *split* possibili condizionatamente alla scelta di A_1. La massimizzazione viene quindi effettuata rispetto ad ogni possibile suddivisione di A nelle due superclassi, per cui lo *split* finale, in base al quale viene effettuata la suddivisione sul nodo t, è $s^*(A_1^*)$. Tale *split* è quello che massimizza la funzione $\Delta i[s^*(A_1), t, A_1]$. Il *twoing* e l'*ordered twoing* operano secondo lo schema tracciato con l'unica differenza che il secondo criterio viene applicato qualora sia ragionevole considerare un ordinamento fra le modalità della variabile dipendente, per cui le superclassi possono essere formate solo rispetto a categorie tra loro contigue.

È stato dimostrato (Breiman *et al.*, 1984) che lo *split* ottimale $s^*(A_1^*)$ qualora si applichi il criterio *twoing* si ottiene massimizzando la seguente funzione:

$$\Phi(s,t) = \frac{p_r p_l}{4} \left[\sum_j |p(j|t_l) - p(j|t_r)| \right]^2 .$$

Conseguentemente il metodo *twoing* non richiede la fissazione d'una misura di impurità dei nodi per la massimizzazione di $\Delta i(s,t)$.

5.2. *La ricerca del sottoalbero ottimale: il* pruning

Come è stato anticipato nell'introduzione di questo capitolo, il tratto più originale della metodologia CART consiste nella proposta d'un metodo per la validazione dell'albero. Tale criterio rappresenta una regola di stop nella procedura di costruzione dell'albero di classificazione. Intuitivamente, un criterio ragionevole di stop è quello di fissare una soglia minima β per il decremento di impurità dell'albero passando da uno stadio a quello successivo, al di sotto la quale la procedura si arresta, cioè:

$$\max_{s \in S} \Delta IMP(s,t) < \beta \tag{7.10}$$

La scelta soggettiva della soglia influenza pesantemente i risultati. Infatti,

— se β è troppo piccolo è probabile ottenere un albero finale profondo (cioè un albero con molte foglie) con conseguenti difficoltà interpretative.

— se β è troppo elevato un nodo t può essere dichiarato terminale — poiché $\Delta IMP(s,t) < \beta$ — escludendo la possibilità che i suoi nodi discendenti ammettano un decremento di impurità $\geq \beta$.

Una considerazione ulteriore riguarda la stima di $R(T)$, dove T è un generico albero di classificazione. La stima per risostituzione $\widehat{R}(T)$ è inversamente proporzionale al numero di foglie dell'albero. L'accuratezza di $\widehat{R}(T)$ decresce al crescere delle dimensioni dell'albero e la scelta della dimensione dell'albero basata esclusivamente su tale stima porta alla selezione di classificatori con numerosi *split* (16). La stima basata sul campione test $\widehat{R}_{ts}(T)$ ha, invece, prima un andamento decrescente e poi crescente all'aumentare del numero di foglie, oltre una certa soglia.

Le considerazioni effettuate in relazione alla scelta del criterio di stop e alla stima del tasso di errata classificazione hanno condotto all'introduzione d'una metodologia di validazione degli alberi detta *pruning* (potatura) le cui fasi possono essere sintetizzate come segue:

---

(16) Tale eventualità può verificarsi poiché la scelta del migliore *split* è ottimale solo in relazione ad ogni singolo stadio.

1) creazione dell'albero massimo T_{max} che si ottiene fissando $\beta = 0$, per cui le foglie sono costituite da casi appartenenti alla stessa classe o al limite da un solo caso;

2) selezione dei sottoalberi che si possono ottenere tagliando T_{max} in determinati punti e stima del tasso di errata classificazione dei diversi sottoalberi mediante uno stimatore appropriato di $R(T)$; tale fase costituisce il nucleo del *pruning*, poiché l'albero viene sfrondato eliminando alcuni rami « secondari »;

3) scelta del sottoalbero che fornisce la migliore stima di $R(T)$.

Il numero di possibili sottoalberi può essere molto elevato anche quando l'albero T_{max} ha un numero limitato di foglie. Al fine di limitare la complessità computazionale legata all'analisi di tutti i possibili sottoalberi, si utilizza una procedura di *pruning* selettivo, cioè un metodo che consente di individuare una sequenza di sottoalberi di dimensione decrescente $T_{max}, T_1, T_2, ..., \{t_1\}$, dove $\{t_1\}$ è l'albero costituito solo dal nodo radice. Ogni sottoalbero appartenente alla sequenza ottimale è il « migliore » rispetto ai sottoalberi appartenenti alla stessa classe, cioè rispetto ai sottoalberi aventi il medesimo numero di foglie. Al fine di individuare la sequenza ottimale si definisce, per ogni albero $T \leq T_{max}$, una misura $R_\alpha(T)$ detta funzione di costo-complessità, cioè

$$R_\alpha(T) = \widehat{R}(T) + \alpha\left|\widetilde{T}\right| \tag{7.11}$$

dove $\left|\widetilde{T}\right|$ è il numero di foglie dell'albero T, $\widehat{R}(T)$ è la stima per risostituzione del tasso di errata classificazione (17) e α è un numero reale non negativo detto parametro di complessità. Tale parametro può essere considerato come una penalità connessa ad alberi di grande dimensione, per cui fra due alberi aventi lo stesso valore di $\widehat{R}(T)$ si seleziona quello con il minore numero di foglie. Fissato il valore del parametro di complessità si ricerca quel sottoalbero $T(\alpha) \leq T_{max}$ tale che

$$R_\alpha[T(\alpha)] = \min_{T \leq T_{max}} R_\alpha(T).$$

Se α è piccolo, la penalità connessa ad un numero elevato di foglie è bassa per cui $T(\alpha)$ sarà complesso. Al crescere di α, $T(\alpha)$ avrà un

---

(17) Come è stato anticipato, l'accuratezza di $R(T)$ decresce al crescere della dimensione dell'albero. Nonostante ciò, $R(T)$ è una misura adeguata per il confronto di alberi aventi la stessa dimensione.

numero sempre inferiore di foglie finché, per un valore di α sufficientemente elevato, il sottoalbero ideale sarà quello formato dal solo nodo radice.

Sebbene α appartenga al campo dei numeri reali, il numero di sottoalberi di $T_{\max}$ è sempre finito, per cui il processo di *pruning* determina una sequenza finita di sottoalberi con un numero di foglie decrescente al crescere di α. Si può dimostrare (Breiman *et al.*, 1984) che $\forall\alpha$ esiste un unico sottoalbero $T \leq T_{\max}$ che minimizza $R_\alpha(T)$, per cui la sequenza ottimale di sottoalberi $T_1, T_2, \ldots, \{t_1\}$, con $T_k = T(\alpha_k)$, $\alpha_1 = 0$, è identificata in modo univoco.

I sottoalberi appartenenti alla sequenza ottimale vengono quindi confrontati utilizzando una stima del tasso di errata classificazione $\tilde{R}(T_k)$. Il sottoalbero ideale T_{k_0} sarà quello per cui la stima del tasso di errata classificazione è minima, cioè:

$$\tilde{R}(T_{k_0}) = \min_k \tilde{R}(T_k). \tag{7.12}$$

La scelta del miglior sottoalbero è chiaramente influenzata dallo stimatore $\tilde{R}(T)$ utilizzato. Infatti, nel caso in cui si utilizzi la stima per risostituzione $\hat{R}(T)$, il sottoalbero selezionato sarà quello più complesso, cioè T_1. È quindi necessario ricorrere a stime più accurate del tasso di errata classificazione rappresentate da $\hat{R}_{ts}(T)$ e da $\hat{R}_{cv}(T)$. Per studiare l'accuratezza d'una stima in termini di *standard error* è necessario definire un modello di probabilità. Si assume quindi che i casi costituenti il campione di partenza S siano fra loro indipendenti e siano tratti dalla distribuzione di probabilità $P(\cdot)$ definita nello spazio $\mathbf{X} \times A$, dove $\mathbf{X}$ è lo spazio delle variabili esplicative, mentre A è l'insieme delle modalità assumibili dalla variabile dipendente. La probabilità di classificare erroneamente un caso di classe j nella classe j' utilizzando d è dato da:

$$Q(j'|j) = P[d(\mathbf{X}) = j'|Y = j]. \tag{7.13}$$

Se si indica con $C(j'|j)$ il costo che si sostiene classificando erroneamente un caso di classe j nella classe j', il costo atteso di errata classificazione dei casi di classe j sarà

$$R(j) = \sum_{j'} C(j'|j)Q(j'|j). \tag{7.14}$$

Il costo di errata classificazione connesso al classificatore d sarà invece

$$R(d) = \sum_j R(j)\pi(j) \tag{7.15}$$

dove $\pi(j)$ è la probabilità a priori che un caso venga classificato nella classe j.

Gli indicatori (7.13), (7.14), (7.15) possono essere stimati attraverso il metodo del campione test o mediante la *cross-validation*. In entrambi i casi, l'idea centrale è che $Q(j'|j)$ possa essere stimato in base alle frequenze delle osservazioni classificate erroneamente. Inoltre, è possibile calcolare gli errori standard delle stime ipotizzando che la stima di $Q(j'|j)$ si distribuisca secondo una variabile aleatoria binomiale.

5.2.1. *Stime mediante un campione test*

L'albero $T_{\max}$ viene costruito utilizzando solo il campione d'apprendimento S_1 e viene «potato» al fine di ottenere la sequenza ottimale $T_1 > T_2 > \ldots > \{t_1\}$. Successivamente, i casi appartenenti al campione test S_2 vengono classificati utilizzando ogni albero T_k appartenente alla sequenza ottimale. Poiché la vera classe di appartenenza dei casi in S_2 è nota, è possibile calcolare il costo di errata classificazione di ogni T_k al fine di stimare $R_{ts}(T_k)$.

Scendendo nel dettaglio, si indichi con $n_j^{(2)}$ il numero di unità statistiche del campione S_2 appartenenti alla classe j. Per ogni sottoalbero T_k si indichi con $n_{j'j}^{(2)}$ il numero di osservazioni di classe j del campione test classificate nella classe j'. La stima di (7.13) sarà data dalla proporzione di osservazioni del campione test appartenenti alla classe j, ma erroneamente classificate nella classe j', cioè:

$$\widehat{Q}_{ts}(j'|j) = \frac{n_{j'j}^{(2)}}{n_j^{(2)}}. \tag{7.16}$$

Utilizzando la (7.14) e la (7.15) si stimano i corrispondenti indicatori come segue:

$$\widehat{R}_{ts}(j) = \sum_{j'} C(j'|j)Q_{ts}(j'|j) \tag{7.17}$$

$$\widehat{R}_{ts}(T) = \sum_j \widehat{R}_{ts}(j)\pi(j). \tag{7.18}$$

Le probabilità a priori $\pi(j)$ che un caso appartenga alla classe j possono essere stimate attraverso il rapporto $n_j^{(2)}/n^{(2)}$ dove $n^{(2)}$ è la numerosità del campione test. In tale caso la (7.18) diventa:

$$\widehat{R}_{ts}(T) = \sum_j \sum_{j'} C(j'|j) \frac{n_{j'j}^{(2)}}{n^{(2)}}. \tag{7.19}$$

Nel caso in cui i costi di errata classificazione siano unitari (18), $\widehat{R}_{ts}(j)$ è la proporzione di casi di classe j erroneamente classificati. Se, inoltre, le probabilità a priori $\pi(j)$ vengono stimate attraverso i dati, $\widehat{R}_{ts}(T)$ è la proporzione totale di casi del campione test classificati erroneamente da T.

L'albero ideale T_{k_0} all'interno della sequenza ottimale verrà quindi selezionato in base alla seguente regola:

$$\widehat{R}_{ts}(T_{k_0}) = \min_k \widehat{R}_{ts}(T_k) \tag{7.20}$$

5.2.2. *Stime mediante* cross-validation

Come è stato precedentemente definito, nella *V-fold cross-validation* il campione originale S viene diviso casualmente in V sottoinsiemi, S_v, $v = 1, 2, \ldots, V$, ognuno contenente (approssimativamente) lo stesso numero di casi. Quindi, per ogni v, S_v viene considerato come un campione test e $S^{(v)} = S - S_v$ rappresenta il campione di apprendimento. Nel caso della *cross-validation* si costruisce un albero massimale $T_{\max}^{(v)}$ per ogni v utilizzando solo le osservazioni appartenenti a $S^{(v)}$. Quindi, per ogni v e per ogni valore del parametro di complessità α si costruisce il corrispondente albero ottimale $T^{(v)}(\alpha)$ utilizzando la funzione di costo-complessità (7.11).

Dopo aver creato gli alberi ottimali per ogni v e per ogni α, gli errori di classificazione vengono calcolati utilizzando i campioni test S_v.

---

(18) In tale caso le conseguenze di un errore di classificazione sono ritenute indipendenti dalla vera classe di appartenenza. Ad esempio, nel caso di una variabile dipendente dicotomica che assume le modalità 0 e 1, l'errore che si commette classificando una unità appartenente alla classe 0 nella classe 1 ha la stessa gravità dell'errore inverso.

Fissato un valore del parametro di complessità α, per ogni valore v, j e j' si definisce $n_{j'j}^{(v)}$ come il numero di casi di classe j (appartenenti al campione S_v) attribuiti erroneamente alla classe j', per cui il numero totale di casi di classe j (appartenenti ad un qualsiasi campione test) attribuiti alla classe j' sarà:

$$n_{j'j} = \sum_v n_{j'j}^{(v)}$$

Poiché ogni caso in S cade in un solo campione test S_v, il numero totale di casi di classe j appartenenti a tutti i campioni test è pari a n_j. In conseguenza, per un dato valore di α le stime di errata classificazione dell'albero saranno:

$$\hat{Q}_{cv}(j'|j) = \frac{n_{j'j}}{n_j} \tag{7.21}$$

$$\hat{R}_{cv}(j) = \sum_{j'} C(j'|j)\hat{Q}_{cv}(j'|j) \tag{7.22}$$

$$\hat{R}_{cv}(T(\alpha)) = \sum_j \hat{R}_{cv}(j)\pi(j). \tag{7.23}$$

Nel caso in cui le probabilità siano stimate attraverso i dati, per cui $\pi(j) = n_j/n$, la (7.23) diventa:

$$\hat{R}_{cv}(T(\alpha)) = \frac{1}{n}\sum_j \sum_{j'} C(j'|j)n_{j'j}, \tag{7.24}$$

e se i costi sono unitari, la (7.24) rappresenta semplicemente la proporzione di casi del campione test che sono classificati erroneamente. Fino ad ora si è ipotizzato che α sia fisso. Come è già stato sottolineato, anche se α varia nel continuo, gli alberi a costo-complessità minimale sono uguali a T_k per $\alpha_k \leq \alpha < \alpha_{k+1}$. Si consideri $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$, per cui α'_k è la media geometrica degli estremi dell'intervallo per il quale $T(\alpha) = T_k$. Quindi, si ponga:

$$\hat{R}_{cv}(T_k) = \hat{R}_{cv}[T(\alpha'_k)] \tag{7.25}$$

dove $\hat{R}_{cv}[T(\alpha'_k)]$ viene calcolato secondo la (7.23). $\hat{R}_{cv}(T_k)$ è la stima che si ottiene classificando i campioni S_v mediante gli alberi $T^{(v)}(\alpha'_k)$.

La regola per la scelta dell'albero ideale T_{k_0} sarà quindi:

$$\widehat{R}_{cv}(T_{k_0}) = \min_k \widehat{R}_{cv}(T_k). \tag{7.26}$$

5.2.3. *La regola: «una volta lo* standard error*»*

Le regole (7.20) e (7.26) si basano su stime del tasso di errata classificazione la cui precisione può essere valutata mediante la stima dello *standard error* corrispondente (Breiman *et al.*, 1984, pp. 78-80). Attraverso numerosi esempi e studi di simulazione è stato osservato che, rappresentando graficamente le stime del tasso di errata classificazione $\widetilde{R}(T)$ (calcolate con il metodo del campione test o con la *cross-validation*) rispetto al numero di foglie dell'albero corrispondente, si ottiene un andamento che è piuttosto piatto all'interno della regione delimitata dagli estremi $\widetilde{R}(T) - SE(\widetilde{R}(T))$ e $\widetilde{R}(T) + SE(\widetilde{R}(T))$, dove $SE(\widetilde{R}(T))$ indica lo *standard error* di $\widetilde{R}(T)$.

Se si definisce T_{k_0} attraverso la seguente procedura:

$$\widetilde{R}(T_{k_0}) = \min_k \widetilde{R}(T_k),$$

secondo la regola «una volta lo *standard error*» l'albero ideale è T_{k_1}, dove k_1 è il massimo valore di k che soddisfa la seguente disuguaglianza:

$$\widetilde{R}(T_{k_1}) \leq \widetilde{R}(T_{k_0}) + SE\left[\widetilde{R}(T_{k_0})\right]. \tag{7.27}$$

5.3. *Un esempio*

A titolo d'esempio prendiamo in considerazione 241 aziende appartenenti al comparto tessile della provincia di Prato (19). Per ogni azienda si dispone di 25 indici di bilancio riferiti all'anno 1996 (ROI; ROS; valore aggiunto/fatturato; valore aggiunto al netto delle spese generali/fatturato; circolante/totale attività; risultato operativo/oneri finanziari; oneri finanziari/fatturato; fatturato/circolante; fatturato/totale attività; capitale netto/fatturato; utile d'esercizio/fatturato; *cash flow*/fatturato; *cash flow*/totale attività; utile d'esercizio/totale attività; *cash flow*/debiti; debiti/totale attività; capitale netto/debiti; rimanenze/fat-

(19) Dati forniti dalla Provincia di Prato - SIEL, Centrale dei Bilanci. Si ringrazia vivamente la provincia di Prato per avere consentito l'utilizzazione dei dati in questa sede.

turato; consumi di materie prime/costo della produzione; spese generali/costo della produzione; personale/costi della produzione; ammortamenti/costi della produzione; fondo trattamento di fine rapporto/totale attività; fondo trattamento di fine rapporto/debiti; fondo trattamento di fine rapporto/cicolante al netto delle rimanenze) e del settore di attività in cui essa opera che può assumere 11 modalità (ad es.: filature a pettine, maglifici, tessiture, etc.). In epoca successiva (1998), si è rilevato che 200 aziende erano ancora operanti e pertanto sono state classificate come «sane», mentre le restanti 41 erano state dichiarate fallite.

La lettura dei bilanci (se la compilazione è stata effettuata correttamente) consente di ricostruire la situazione economico-finanziaria di un'azienda. È ragionevole ipotizzare che le aziende sane abbiano indicatori di bilancio diversi da quelli delle aziende fallite. Partendo da tale constatazione si costruisce nel seguito un albero di classificazione secondo la metodologia CART utilizzando la situazione delle aziende (indicata con Y) come variabile dipendente che si manifesta secondo le modalità sana (codice 0) e fallita (codice 1). I predittori sono rappresentati da 25 indici di bilancio (X_s, $s = 1, 2, ..., 25$) e dal settore di attività (SETTORE con modalità A, B, ..., M).

Nelle prime due fasi della procedura di costruzione dell'albero con *AnswerTree* si procede come nell'applicazione di CHAID (par. 4.4) ponendo la variabile Y nel riquadro riservato alla variabile *target*. Successivamente si seleziona la sequenza *analysis-growing criteria*. Compare una maschera in cui devono essere specificate le opzioni relative alle regole di arresto (*stopping rules*), la misura di impurità e la regola di *pruning*. Fissiamo la profondità massima dell'albero pari a 10 e la numerosità minima del nodo padre e del nodo figlio pari, rispettivamente, a 2 e 1. Numerosità basse dei nodi garantiscono l'ottenimento del T_{max} che verrà in seguito «potato». Il cambiamento minimo di impurità (*minimum change in impurity*) indica la riduzione di impurità d'un nodo al di sotto la quale la procedura si arresta, perché il miglioramento della classificazione è trascurabile rispetto all'aumento della complessità dell'albero. Poniamo tale parametro pari a 0.0001 (valore di *default*). Utilizziamo quale indice di impurità l'indice di Gini e selezioniamo la regola «una volta lo *standard error*» per la procedura di *pruning*.

Dopo aver definito i criteri di crescita dell'albero, dividiamo il campione ponendo il 50% delle osservazioni nel *learning sample* ed il 50% rimanente nel *testing sample* per evitare che un campione test di numerosità troppo bassa (per es. pari a 1/3) comprenda poche aziende fallite.

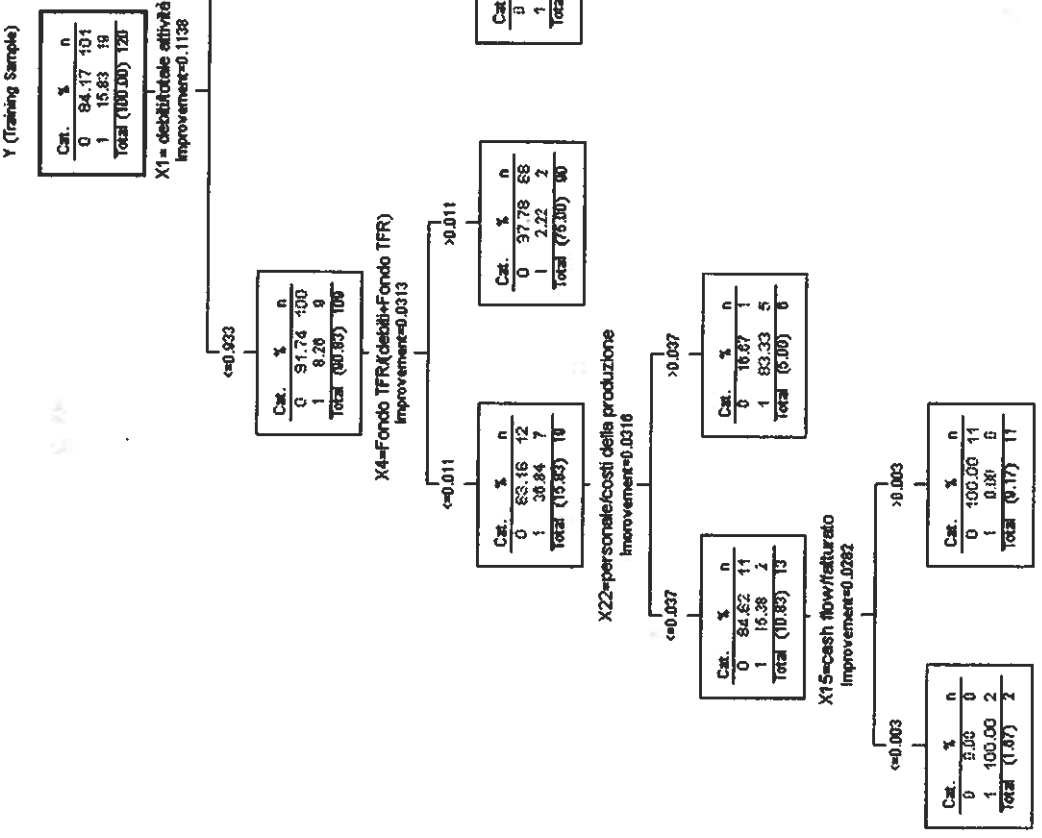


FIG. 7.3. *Albero di classificazione per 241 aziende ottenuto con la metodologia CART. Il codice 0 indica le aziende sane, il codice 1 indica le aziende fallite.*

Selezionando la sequenza *tree-grow tree and prune* si ottiene l'albero (costruito sul *training sample*) riportato nella fig. 7.3.

Il campione di apprendimento comprende 120 aziende (su 241) 19 delle quali sono fallite. Ad ogni *split* (s) viene indicata la misura del decremento di impurità (*improvement*) calcolato secondo la formula (7.7). Le variabili selezionate per la costruzione dell'albero sono X_1 (debiti/totale attività), X_4 (fondo di trattamento fine rapporto/(fondo trattamento fine rapporto+debiti)), X_5 (ROI), X_{15} (*cash flow*/fatturato) e X_{22} (spese per il personale/costi della produzione). Lo *split* che determina la maggiore riduzione di impurità è quello costruito su X_1. Infatti, ponendo come valore soglia 0.933 vengono isolate 11 aziende con elevato indebitamento ($X_1 > 0.933$) 10 delle quali sono fallite. Suddividendo ulteriormente attraverso la X_5 (valore di soglia 0.123) si identifica una foglia che comprende 10 aziende fallite e una foglia con una sola azienda sana. Fra le aziende per le quali $X_1 \leq 0.933$ vengono isolate 90 aziende con $X_4 > 0.011$. Il 97.78% di queste aziende è composto da aziende sane. Analizzando l'albero globalmente si deduce che le aziende fallite hanno una struttura finanziaria caratterizzata da forte indebitamento (X_1 elevato), oppure da una forte incidenza del costo del personale sui costi di produzione (X_{22}) e da un *cash flow* basso rispetto al fatturato (X_{15}). Ad ogni foglia viene assegnata la classe con frequenza superiore. Ad esempio, nella foglia che raccoglie le aziende per le quali $X_1 \leq 0.933$ e $X_4 > 0.011$ 97.78% sono sane, mentre 2.2% sono fallite per cui le nuove aziende che verranno classificate secondo questa regola verranno assegnate alla classe « sana ». Utilizzando la regola di classificazione della fig. 7.3 per suddividere le aziende del campione test, 91 aziende sane e 16 aziende fallite vengono classificate correttamente, 6 aziende fallite vengono erroneamente classificate come sane, mentre l'errore opposto (aziende sane classificate come fallite) si verifica in 8 casi (tab. 7.2).

TAB. 7.2. *Confronto fra modalità reale e modalità prevista dalla regola di classificazione applicata al campione test composto da 121 aziende.*

		Categoria reale	
		0	1	**totale**
Categoria prevista	0	91	6	**97**
	1	8	16	**24**
	totale	99	22	**121**

Il tasso di errata classificazione globale calcolato sul *testing sample* è $\widehat{R}_{ts} = 0.115$. Utilizzando la cross-validation con $V = 10$ si ottiene un tasso di errata classificazione $\widehat{R}_{cv} = 0.145$. Poiché la frequenza relativa delle aziende fallite (0.17) è sensibilmente inferiore a quella delle aziende sane (0.83) i campioni test che si formano nella *cross-validation* con $V = 10$ hanno una bassa probabilità di comprendere aziende fallite. Ne consegue che è più probabile classificare erroneamente un'azienda sana rispetto all'errore opposto. Perciò si ritiene che in questo caso la stima più attendibile dell'errore di classificazione sia fornita da $\widehat{R}_{ts}$, poiché non vale l'assunzione di stabilità sulla quale poggia il calcolo di $\widehat{R}_{cv}$ (ved. par. 3.5).

L'analisi precedente è stata condotta su 241 aziende per le quali si conosceva la situazione effettiva (sana o fallita). La regola di decisione che è stata definita è però utilizzabile anche in sede previsiva, con riferimento ad altre aziende non comprese nel campione esaminato. Queste informazioni possono essere estremamente utili, ad esempio, per un istituto di credito che, in base all'analisi di segmentazione gerarchica, concederà più facilmente un finanziamento ad un'azienda se viene classificata « sana ». A tale scopo potrebbe essere più pericoloso (in termini di perdita del credito concesso) classificare erroneamente un'azienda a rischio di fallimento come azienda sana. Per rendere meno probabile tale eventualità l'albero potrebbe essere costruito variando i costi di errata classificazione, ponendo, ad esempio, il costo dell'errore più grave (azienda fallita prevista come azienda sana) pari al doppio (o ad un multiplo) rispetto al costo dell'errore opposto.

5.4. *Alberi di regressione*

Fino ad ora è stato ipotizzato che la variabile dipendente fosse qualitativa. Nel caso in cui la variabile risposta sia quantitativa le strutture generate da CART vengono denominate « alberi di regressione ».

L'approccio alla costruzione degli alberi di regressione è leggermente più semplice rispetto a quello degli alberi di classificazione. Infatti, nella fase di crescita e in quella di potatura d'un albero di regressione viene utilizzata la stessa misura di impurità. Inoltre, quando la variabile dipendente è quantitativa i casi vengono pesati tutti nello stesso modo senza l'impiego di probabilità a priori.

Si consideri, quindi, una variabile dipendente Y che assume valori nel campo dei numeri reali e p variabili esplicative, quantitative o qualitative, $X_1, X_2, \ldots, X_p$ rilevate su n unità statistiche. Sia $\mathbf{x}_i$ il vettore con-

tenente tutte le informazioni per l'i-esimo caso. L'obiettivo d'un albero di regressione è quello di costruire una funzione $d(\mathbf{x})$ sullo spazio $\mathbf{X}$ delle variabili esplicative, che assuma valori reali. Tale funzione viene definita *regola di previsione* o *previsore.*

Come negli alberi di classificazione, per la costruzione di $d(\mathbf{x})$, lo spazio $\mathbf{X}$ viene suddiviso attraverso una sequenza di *split* binari fino a raggiungere un insieme di nodi terminali. In ogni nodo terminale t il valore previsto per la variabile dipendente $y(t)$ è costante. La costruzione d'una regola di previsione gerarchica avviene attraverso le seguenti fasi:

1) scelta d'un criterio per la selezione d'uno *split* ad ogni nodo intermedio;

2) fissazione d'una regola di stop per l'individuazione dei nodi terminali;

3) costruzione d'una procedura per l'assegnazione di un valore $y(t)$ ad ogni nodo terminale t.

Per definire tali fasi è necessario fissare un criterio di accuratezza della regola di previsione. A tale scopo si utilizza in genere l'errore quadratico medio $R(d)$, del previsore d che può essere stimato secondo diversi criteri (20). Se si utilizza la stima per risostituzione avremo:

$$\hat{R}(d) = \frac{1}{n}\sum_{i=1}^{n} [y_i - d(\mathbf{x}_i)]^2. \tag{7.28}$$

La stima basata sul campione test si ottiene suddividendo casualmente il campione S in due sottocampioni S_1 e S_2. Quindi il previsore viene costruito su S_1 mentre la stima dell'errore quadratico medio sarà calcolata su S_2, cioè:

$$\hat{R}_{ts}(d) = \frac{1}{n_2}\sum_{i \in S_2} [y_i - d(\mathbf{x}_i)]^2. \tag{7.29}$$

Se dividiamo S in V sottoinsiemi $S_1, S_2, \ldots, S_V$ e costruiamo il previsore $d^{(v)}(x)$ su $S - S_v$, con $v = 1, 2, \ldots, V$, la stima *cross-validation* sarà:

---

(20) Per uniformità di notazione, è stata utilizzata la medesima simbologia impiegata per il tasso di errata classificazione negli alberi di classificazione.

$$\widehat{R}_{cv}(d) = \frac{1}{n}\sum_{v=1}^{V}\sum_{i\in S_v}\left[y_i - d^{(v)}(\mathbf{x}_i)\right]^2. \tag{7.30}$$

Rispetto al tasso di errata classificazione calcolato per gli alberi di classificazione, l'errore quadratico medio è influenzato dalla scala in cui è espressa la variabile dipendente. Per rimuovere tale effetto, l'errore quadratico medio viene diviso per la varianza di Y, la quale è stimata come segue:

$$\widehat{R}(\bar{y}) = \frac{1}{n}\sum_{i}(y_i - \bar{y})^2 \tag{7.31}$$

dove $\bar{y}$ è la media aritmetica delle realizzazioni di Y.

Le stime dell'errore quadratico medio relativo ($RE(d)$) per risostituzione, del campione test e di cross-validation saranno, rispettivamente (21),

$$\widehat{RE}(d) = \widehat{R}(d)/\widehat{R}(\bar{y}), \tag{7.32}$$

$$\widehat{RE}_{ts}(d) = \widehat{R}_{ts}(d)/\widehat{R}_{ts}(\bar{y}), \tag{7.33}$$

$$\widehat{RE}_{cv}(d) = \widehat{R}_{cv}(d)/\widehat{R}(\bar{y}). \tag{7.34}$$

Se utilizziamo la stima di $R(d)$ fornita dalla (7.28), il valore di $y(t)$ che minimizza tale stima è la media aritmetica degli y_i relativi a tutti i casi che cadono in t, cioè:

$$\bar{y}(t) = \frac{1}{n(t)}\sum_{i\in t} y_i. \tag{7.35}$$

La dimostrazione di tale proposizione deriva dalle proprietà della media aritmetica (Vol. I, p. 54).

Quindi il problema dell'assegnazione d'un valore ad ogni nodo viene risolto sostituendo ai valori presenti nel nodo la loro media arit-

---

(21) $RE(d)$ è sempre non negativo ed è solitamente minore di 1. Infatti, i previsori ragionevoli di Y saranno più accurati del suo valore atteso μ, ma può verificarsi che alcuni previsori particolarmente inefficaci generino un $RE(d) \geq 1$.

metica, che rappresenta la migliore previsione qualora si scelga la stima per risostituzione di $R(d)$ come misura di accuratezza del previsore.

Senza perdita di generalità possiamo ora sostituire la notazione $R(d)$ con $R(T)$ dove T è un generico albero di regressione e come tale rappresenta un previsore. Se la (7.35) rappresenta la previsione di Y per il nodo t, possiamo scrivere:

$$\widehat{R}(T) = \frac{1}{n} \sum_{t \in \widetilde{T}} \sum_{i \in t} [y_i - \bar{y}(t)]^2. \tag{7.36}$$

Il migliore *split* s^* di un generico nodo t è quello appartenente all'insieme Θ che determina il maggior decremento di (7.36). Per ogni *split* s di t in t_l e t_r, sia:

$$\Delta\widehat{R}(s,t) = \widehat{R}(t) - \widehat{R}(t_l) - \widehat{R}(t_r).$$

Quindi il migliore *split* sarà quello per cui:

$$\Delta\widehat{R}(s^*,t) = \max_{s \in \Theta} \Delta\widehat{R}(s,t). \tag{7.37}$$

Perciò un albero di regressione viene costruito suddividendo iterativamente i nodi al fine di produrre il massimo decremento di $\widehat{R}(T)$. La stima per risostituzione nell'ambito della regressione individua quella soglia di suddivisione dello spazio delle variabili esplicative che separa in maniera più efficace i valori elevati della variabile dipendente da quelli bassi.

La fase finale della costruzione d'un albero di regressione consiste nella individuazione dei nodi terminali. A tal fine non può essere utilizzata la stima per risostituzione perché essa, come nel caso degli alberi di classificazione, ha un andamento monotono al crescere della dimensione dell'albero.

Anche per gli alberi di regressione si utilizza la procedura di *pruning*. Si costruisce l'albero $T_{\max}$, cioè l'albero che si ottiene suddividendo in modo iterativo i nodi intermedi finché essi contengono gli stessi valori o la loro numerosità raggiunge una soglia minima. La creazione della sequenza ottimale avviene utilizzando la seguente misura di errore-complessità:

$$R_\alpha(T) = \widehat{R}(T) + \alpha\left|\widetilde{T}\right|. \tag{7.38}$$

Il risultato è una sequenza decrescente di alberi $T_1 > T_2 > \ldots > \{t_1\}$ con $T_1 \leq T_{\max}$ e una corrispondente sequenza di parametri di complessità $0 = \alpha_1 < \alpha_2 < \ldots$, tale che per $\alpha_k \leq \alpha < \alpha_{k+1}$, T_k è il più piccolo sottoalbero di $T_{\max}$ che minimizza $R_\alpha(T)$. La scelta del miglior sottoalbero all'interno della sequenza ottimale avviene, come per gli alberi di classificazione, utilizzando la stima del campione test $\widehat{R}_{ts}(T_k)$ o la stima *cross-validation* $\widehat{R}_{cv}(T_k)$ dell'errore di previsione, che sono, rispettivamente:

$$\widehat{R}_{ts}(T_k) = \frac{1}{n_2} \sum_{i \in S_2} [y_i - d_k(\mathbf{x}_i)]^2 \tag{7.39}$$

dove $d_k(\mathbf{x})$ rappresenta il previsore corrispondente all'albero T_k, e

$$\widehat{R}_{cv}(T_k) = \frac{1}{n} \sum_{v=1}^{V} \sum_{i \in S_v} \left[y_i - d_k^{(v)}(\mathbf{x}_i) \right]^2 \tag{7.40}$$

dove $d_k^{(v)}(\mathbf{x})$ rappresenta il previsore corrispondente all'albero $T^{(v)}(\alpha'_k)$, con $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$.

Se, per esempio, si stima l'errore quadratico medio di previsione mediante il metodo della *cross-validation* utilizzando la regola «una volta lo *standard error*», l'albero T_k selezionato sarà il più piccolo albero tale che:

$$\widehat{R}_{cv}(T_k) \leq \widehat{R}_{cv}(T_{k_0}) + SE \tag{7.41}$$

dove $\widehat{R}_{cv}(T_{k_0}) = \min_k \widehat{R}_{cv}(T_k)$ (22).

6. *Cenni alla metodologia QUEST*

I metodi di partizione ricorsiva che derivano dalla metodologia CART vengono definiti metodi esaustivi, poiché la scelta del migliore *split* avviene analizzando tutte le possibili suddivisioni che si possono creare su ogni variabile esplicativa. Come è stato sottolineato nel par. 3.1, il numero di possibili dicotomizzazioni (e quindi di *split*) è molto

(22) La procedura di *pruning* qui descritta è quella proposta da Breiman *et al.* (1984). Per altre procedure di dimensionamento ottimale degli alberi, si veda la rassegna di Mingers (1989).

elevato quando le variabili sono continue o nominali con un numero *m* elevato di modalità. Più precisamente, la complessità computazionale dei metodi esaustivi cresce linearmente nel caso di variabili quantitative con *n* distinti valori ed esponenzialmente nel caso di variabili qualitative nominali con *m* modalità. Ciò determina tempi lunghi di elaborazione di *data set* complessi, anche utilizzando calcolatori con caratteristiche tecniche avanzate (Lim *et al.*, 1998).

Inoltre, è stato dimostrato, attraverso studi di simulazione, che i metodi esaustivi presentano una distorsione nella selezione delle variabili, poiché tendono a selezionare i predittori con un numero elevato di *split* (Loh e Shih, 1997).

Il metodo QUEST (*Quick, Unbiased, Efficient, Statistical Tree*; Loh e Shih, 1997) è un perfezionamento del metodo FACT (*Fast Algorithm for Classification Trees*; Loh e Vanichsetakul, 1988) ed è stato introdotto in letteratura per superare i problemi relativi ai metodi esaustivi. Tale algoritmo impiega però metodi statistici che non sono illustrati in questo volume e precisamente l'analisi della varianza (ANOVA) e l'analisi discriminante quadratica (QDA). Pertanto ci limitiamo a fornire solo alcuni cenni sulla tecnica QUEST, rinviando per approfondimenti al lavoro originale di Loh e Shih (1997).

6.1. *I passi della procedura*

Come si noterà in questo paragrafo il tratto caratterizzante della procedura, che la differenzia dai metodi esaustivi, consiste nella separazione della fase di selezione delle variabili da quella di selezione dello *split*.

— *La selezione delle variabili.* Per evitare la distorsione dei metodi esaustivi, la selezione delle variabili viene eseguita mediante l'impiego di test per la verifica d'ipotesi. In particolare nel caso di variabili esplicative qualitative viene eseguito un test χ^2 per verificare l'indipendenza rispetto alle modalità della variabile dipendente. Nel caso di variabili esplicative quantitative si effettua un'analisi della varianza (ANOVA) per verificare la significatività delle differenze fra le medie calcolate rispetto alle classi della variabile dipendente (23).

---

(23) Per un'introduzione all'analisi della varianza si rinvia a Cicchitelli (1994) e per un approfondimento a Casella and Berger (1990).

— *La selezione del punto di suddivisione.* La fase di selezione dello *split* ottimale viene eseguita attraverso l'applicazione d'una forma modificata dell'analisi discriminante quadratica (QDA, McLachlan, 1992) considerando la variabile selezionata nella fase precedente come variabile esplicativa.

La QDA è applicabile solo qualora le variabili esplicative siano quantitative per cui, se la variabile selezionata è qualitativa, è necessario trasformarla al fine di sostituire alle modalità nominali dei codici numerici.

— *Il criterio di arresto.* La procedura QUEST prevede la possibilità di arrestare la crescita dell'albero utilizzando il metodo di *pruning* previsto per la metodologia CART. Ovviamente si possono utilizzare regole di stop più elementari quali la profondità massima dell'albero (numero massimo di stadi di suddivisione ammessi) o la numerosità minima delle foglie. Non è possibile utilizzare il criterio legato alla misura di impurità perché essa non viene definita.

Il modulo *AnswerTree* prevede la procedura QUEST, ma per la sua utilizzazione richiede la selezione d'un insieme di opzioni legate all'analisi della varianza e all'analisi discriminante quadratica.

7. *Conclusioni e approfondimenti*

Nel presente capitolo sono stati presentati alcuni metodi di segmentazione «classici» con particolare riferimento alle procedure disponibili all'interno del *package* statistico SPSS (modulo *AnswerTree*). La gamma di applicazioni rese possibili da tali procedure è piuttosto vasta. Nella tab. 7.3 vengono riportate le più rilevanti opzioni previste dalle tre metodologie disponibili in *AnswerTree.*

TAB. 7.3. *Principali caratteristiche degli algoritmi di segmentazione.*

Algoritmo	Segmentazione	Variabile dipendente	Predittori
CHAID	Multipla	Qualitativa	Qualitativi
CART	Binaria	Qualitativa e quantitativa	Qualitativi e quantitativi
QUEST	Binaria	Qualitativa	Qualitativi e quantitativi

Naturalmente la rassegna presentata in questo capitolo non pretende di essere esaustiva, poiché esistono molte altre tecniche di segmentazione basate sugli alberi (Pallara, 1992; Lim *et al.*, 1998).

Nella letteratura statistica nazionale ed internazionale recente sono comparsi numerosi articoli collegati alla segmentazione gerarchica in cui sono contenute integrazioni e modifiche dei procedimenti esistenti, nuove procedure o applicazioni originali di tecniche conosciute. In particolare, Capiluppi *et al.* (1999) hanno implementato un *software* statistico per la costruzione di alberi binari o ternari con variabile dipendente e variabili esplicative sia qualitative, sia quantitative. Siciliano e Mola (1998) suggeriscono l'utilizzo della rappresentazione fattoriale dell'analisi delle corrispondenze asimmetrica per la costruzione di alberi di classificazione ternari. Esposito *et al.* (1998) hanno introdotto un nuovo algoritmo che, a partire dalla distinzione fra due tipi di nodi, combina le possibilità offerte dagli alberi di regressione con quelle degli alberi di classificazione. Come è stato accennato, CART prevede la possibilità di definire gli *split* su combinazioni lineari di variabili. Tale selezione è basata sulla minimizzazione di una misura di impurità. Broedley e Utgoff (1995) propongono una procedura chiamata «albero multivariato» in cui gli *split* su combinazioni lineari di variabili vengono individuati mediante test statistici. Le proprietà dei modelli basati sulla segmentazione gerarchica vengono riassunte efficacemente da White e Liu (1997). Nello stesso lavoro vengono prese in considerazione le analogie degli alberi di classificazione con altre metologie quali l'analisi discriminante non parametrica e i modelli logit.

Fra le numerose applicazioni degli alberi di regressione in ambito economico accenniamo al lavoro di Benedetti (1997), che utilizza gli alberi di regressione per la stima disaggregata su base comunale e sub-comunale di alcune variabili economiche (redditi e consumi delle famiglie suddivisi in capitoli di spesa).

RIFERIMENTI BIBLIOGRAFICI

BELSON, W. A. (1959), Matching and prediction on the principle of biological classification, *Applied Statistics*, vol. 8, pp. 65-75.

BENEDETTI, R. (1997), Reddito e consumi: una soluzione al problema della disaggregazione territoriale basata sugli alberi di regressione, *Quaderni di Statistica e Matematica dell'Università di Trento*, vol. XIX, pp. 4-37.

BIGGS, D., DE VILLE, B. and SUEN, E. (1991), A method of choosing multiway partitions for classification and decision trees, *Journal of Applied Statistics*, vol. 18, pp. 49-62.

BISHOP, Y. M., FIENBERG, S. E. and HOLLAND, P. W. (1975), *Discrete Multivariate Analysis*, MIT press, Cambridge.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. and STONE, C. J. (1984), *Classification and Regression Trees*, Wadsworth, Belmont.

BONFERRONI, C. E. (1936), Teoria statistica delle classi e calcolo delle probabilità, *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3-62.

BRODLEY, C. E. and P. E. UTGOFF, P. E. (1995), Multivariate decision trees, *Machine Learning*, 19, pp. 45-77.

CAPILUPPI, C., FABBRIS, L. e SCARABELLO, M. (1999), UNAIDED: a PC system for binary and ternary segmentation analysis, in: VICHI M., O. OPITZ (eds.), *Classification and Data Analysis. Theory and Application*, Springer Verlag, Berlino, pp. 367-374.

CASELLA, G. and BERGER, R. L. (1990), *Statistical Inference*, Wadsworth & Brooks-Cole, Pacific Grove.

CENTRALE DEI BILANCI, (a cura della) (1998), Alberi decisionali e algoritmi genetici nell'analisi del rischio d'insolvenza, *Bancaria*, n. 1, pp. 74-82.

CICCHITELLI, G. (1994), *Probabilità e statistica* (Quindicesima ristampa), Maggioli, Rimini.

CLARK, A. L. and PREGIBON, D. (1992), Tree-based models, in: CHAMBERS J. M., T. J. HASTIE (eds.), *Statistical Models in S*, Wadsworth & Brook, Pacific Grove, California, pp. 377-419.

ESPOSITO, F., MALERBA, D. e TAMMA, V. (1998), Efficient data-driven construction of model-trees, *Atti del convegno «New Technology and Techniques for Statistics» NTTS98*, Sorrento, 4-6 Novembre 1998, pp. 163-168.

FABBRIS, L. (1997), *Statistica multivariata. Analisi esplorativa dei dati*, McGraw-Hill Italia, Milano.

GNANADESIKAN, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley, New York.

GROSSI, L. e GANUGI, P. (1999), Variable selection for the classification of firms, *CLADAG-SIS 99*, Roma, pp. 141-144.

HAND, D. J. and HENLEY W. E. (1997), Some developments in statistical credit scoring, in: NAKHAEIZADEH, G. and TAYLOR, C. C. (eds.), *Machine Learning and Statistics. The Interface*, pp. 221-238, Wiley, New York.

HARTIGAN, J. A. and WONG, M. A. (1979), Algorithm 136. A *k*-means clustering algorithm, *Applied Statistics*, vol. 28, pp. 100-108.

HAWKINS, D. M. and KASS, G. V. (1982), Automatic Interaction Detection, in: HAWKINS, D.M. (ed.), *Topics in Applied Multivariate Analysis*, Cambridge University Press, Cambridge, pp. 269-302.

KASS, G. V. (1980), An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, vol. 29, pp. 119-127.

KEPTRA, S. (1996), Non-binary classification trees, *Statistics and Computing*, vol. 6, pp. 231-243.

LIM, T. S., LOH, W. Y. and SHIH, Y. S. (1998), An empirical comparison of decision trees and other classification methods, *Technical Report 979*, University of Wisconsin, Madison.

LOH, W. Y. and SHIH, Y. S. (1997), Split selection methods for classification trees, *Statistica Sinica*, vol. 7, pp. 815-840.

LOH, W. Y. and VANICHSETAKUL, N. (1988), Tree-structured classification via generalized discriminant analysis, *Journal of the American Statistical Association*, vol. 83, pp. 715-728.

MINGERS, J. (1989), An empirical comparison of pruning methods for decision tree induction, *Machine Learning*, vol. 4, pp. 227-243.

MCLACHLAN, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

MOLA, F. e SICILIANO, R. (1997), A fast splitting procedure for classification trees, *Statistics and Computing*, vol. 7, pp. 209-216.

MOLTENI, L. (1993), *L'analisi multivariata nelle ricerche di marketing: applicazioni alla segmentazione della domanda e al mapping multidimensionale*, EGEA, Milano.

MORGAN, J. N. and SONQUIST, J. A. (1963), Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association*, vol. 58, pp. 415-434.

PALLARA, A. (1992), Binary decision trees approach to classification: a review of CART and other methods with some applications to real data, *Statistica Applicata*, vol. 4, pp. 253-286.

RIPLEY, B. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

SICILIANO, R. e MOLA, F. (1998), Ternary classification trees: a factorial approach, in: GREENACRE, M. e BLASIUS, J. (eds.), *Visualization of Categorical Data*, Academic Press, CA, pp. 311-324.

TUTTE, W. T. (1984), *Graph Theory*, Encyclopedia of Mathematics, vol. 21, Addison-Wesley, Menlo Park, California.

WHITE, A. P. and LIU, W. Z. (1997), Statistical properties of tree-based approaches to classification, in: NAKAHEIZADEH, G. and TAYLOR, C. C. (eds.), *Machine Learning and Statistics. The Interface*, Wiley, New York, pp. 23-44.