**POLITECNICO DI TORINO**
ESAMI DI STATO PER L'ABILITAZIONE
ALLA PROFESSIONE DI INGEGNERE
I SESSIONE - ANNO 1998
Ramo: INFORMATICA Tema n. 2

Al San Diego Medical Center (California), quando viene ricoverato un paziente con un attacco cardiaco, vengono misurate 19 variabili durante le prime 24 ore. Esse includono la pressione sanguigna, l'età e altre 17 variabili binarie o continue che riassumono i sintomi clinici considerati come indicatori importanti della condizione del paziente. Il vettore delle misure che si riferiscono ad un paziente è comunemente detto *pattern*.
Successivamente si utilizza un classificatore strutturato ad albero, come quello in fig. 1, per determinare se il paziente è o meno ad alto rischio (F = non alto rischio, G = alto rischio).

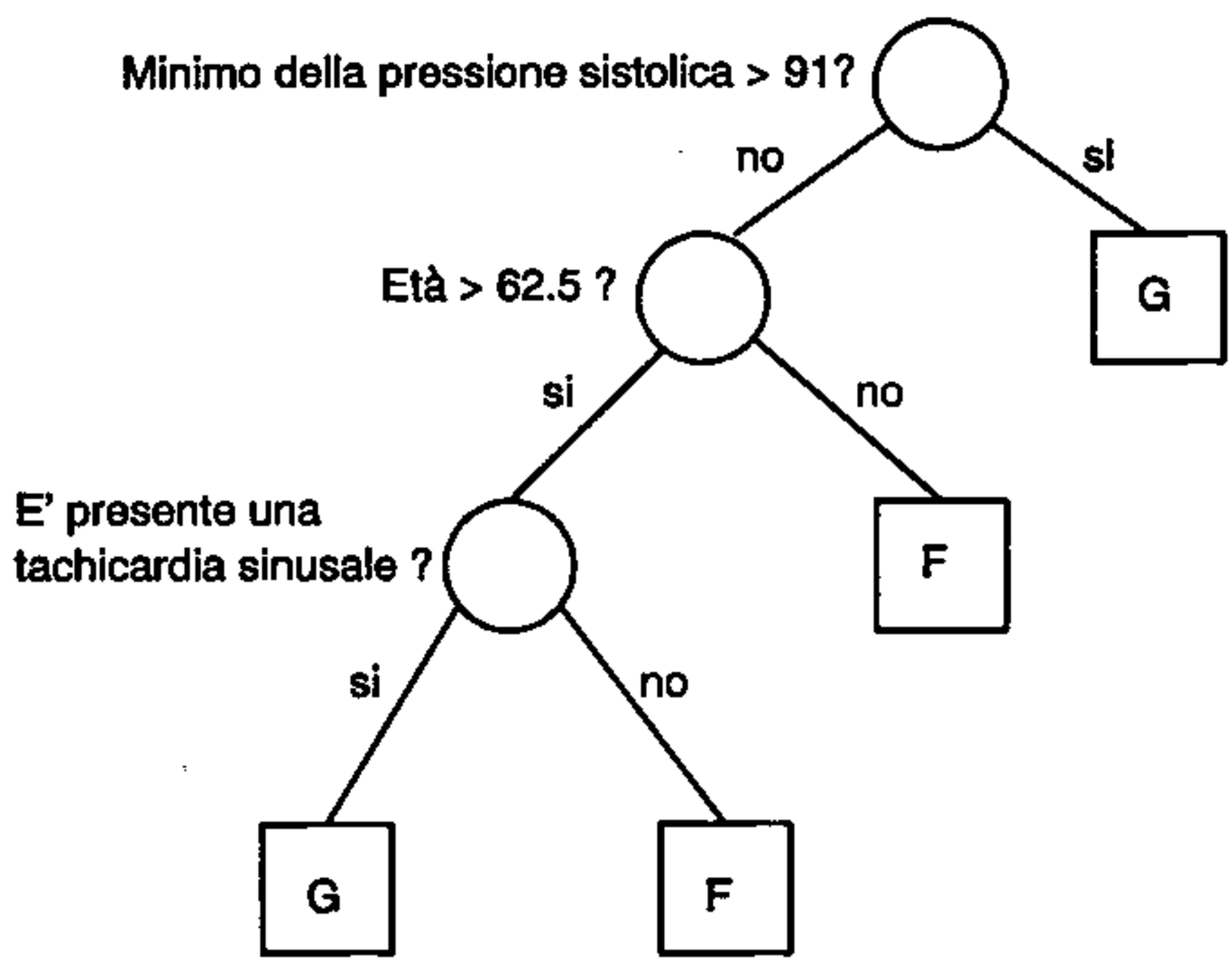


Fig. 1

Il concetto su cui sono basati gli alberi (binari) di classificazione è quello di determinare ad ogni nodo, mediante la verifica di un test, un ordinamento dei dati che porti a "isolare" le singole classi.

Siano dati in un file di tipo testo un gran numero di *pattern* già classificati. I parametri che caratterizzano i pattern (features) possono essere variabili continue (ad esempio, una lunghezza) o categoriali (ad esempio, un colore). Nell'intestazione del file vi è la descrizione completa della struttura dei pattern, come illustrato in seguito.

È richiesto di determinare i parametri di un albero di classificazione secondo le seguenti specifiche.

1. In ogni nodo, la separazione dipenda dal valore di una *singola* variabile.
2. Se la variabile presa in considerazione in un nodo è di tipo continuo, la domanda in un nodo sia nella foma (È $x_m \leq s$ ?), dove $x_m$ è la variabile scelta e $s$ è un valore di soglia opportunamente calcolato.
3. Se la variabile presa in considerazione in un nodo è di tipo categoriale (enumerativo) appartenente ad un set $\{b_1, b_2, ..., b_L\}$, la domanda in un nodo sia nella forma (È $x_m \in S$ ?) dove $S$ è uno dei possibili subset di $\{b_1, b_2, ..., b_L\}$ opportunamente calcolato.
4. Il criterio con cui scegliere la variabile per la separazione (*split*) sia quello di massimizzare il *decremento* di *impurità* (vedi oltre) dell'albero.
5. Il criterio con cui scegliere il blocco dello split sia quello di fissare una soglia minima per il decremento di *impurità* dell'albero.

Per quanto concerne l'*impurità*,

- occorre ideare una funzione nonnegativa che
  a) dia il valore massimo se, in un nodo, sono presenti elementi di tutte le classi nella proporzione in cui le classi sono presenti nel set di addestramento;
  b) dia valore 0 se nel nodo è presente una sola classe.
- concepire il *decremento d'impurità* come l'impurità del nodo di partenza a cui si sottrae l'impurità del nodo di destra e quella del nodo di sinistra, pesate con la porzione di pattern che vanno a cadere nei due rispettivi nodi. In formula: $\Delta i(t) = i(t) - p_L i(t_L) - p_R i(t_R)$, dove $t$ è il nodo padre, $t_L$ e $t_R$ sono i nodo figli sinistro e destro, $p_L$ e $p_R$ le porzioni di pattern nei rispettivi nodi.

Per quanto riguara le variabili continue, è accettabile l'ipotesi che la funzione d'impurità venga calcolata nei soli punti presenti nei pattern d'addestramento e che il comportamento della funzione d'impurità sia di tipo unimodale (è presente un solo minimo: se i pattern sono ordinati in base a quella variabile, si può applicare un metodo dicotomico per individuare il minimo).

Per quanto riguarda le variabili numerabili, è accettato che, in prima approssimazione, si considerino subset composti da un solo elemento.

Una funzione d'impurità accettabile è la seguente:

$$i(t) = -\sum_{1}^{c} p(j/t) \log p(j/t)$$

dove $c$ è il numero di classi e $p(j/t)$ è la proporzione di pattern di classe $j$ presenti nel nodo $t$.

Il file che contiene i pattern d'addestramento è composto da un'intestazione e da un corpo. L'intestazione contiene una descrizione codificata della struttura dei pattern, il corpo contiene i pattern.

Quest'ultimi sono vettori di valori del tipo fissato nell'intestazione, preceduti da un intero che codifica la classe di appartenenza (da 1 a C, se C sono le classi). Non è richiesto di eseguire nessun controllo di validità sul file di addestramento.

Nell'intestazione sono presenti le seguenti informazioni (altre potrebbero essere previste dall'esaminando, se necessarie), una per riga:

*num_pattern*: numero di pattern contenuti nel corpo del file (MAX = 500);
*num_classi*: numero di classi (MAX = 6);
*num_coord*: numero di coordinate (features) in ciascun pattern (MAX = 4);
*F*: simbolo per variabile continua (di tipo floating point);
*N*: simbolo per variabile di tipo numerabile: in questo caso, segue:
*numero*: numero di valori assunti dal tipo numerabile (MAX = 10);
*valore1*: primo valore tra quelli che può assumere il tipo numerabilie;
. . .
*valoreN*: ultimo valore tra quelli che può assumere il tipo numerabile.

Descrivere nel dettaglio ma in modo sintetico le strutture dati e gli algoritmi utilizzati per realizzare il progetto. Il programma può essere codificato in C oppure in PASCAL.
Specificare la forma di output che si intende adottare (lista dei nodi opportunamente contrassegnati, con accanto la domanda, o altro).