Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea in Ingegneria Informatica

Anno accademico 2009/2010

Analisi di Dati Clinici Relativi alla Terapia per l'Epatite C

Relatori:
Ch.mo Prof. Andrea Pietracaprina
Ch.mo Prof. Geppino Pucci

Laureando:
Simone Romano

Correlatori:
Dott.ssa Luisa Cavalletto
Dott.ssa Liliana Chemello

Data di laurea: *26 Ottobre 2010*

Sommario

Questa tesi ha lo scopo di sviluppare dei modelli matematici originali, studiati e validati per l'analisi predittiva della risposta sostenuta (probabilità di guarigione del singolo individuo) alla terapia antivirale per l'*epatite C*. Tali modelli non si limitano alla sola predizione, ma propongono come possa variare l'efficacia terapeutica in base alle scelte compiute dal medico. L'importanza di questa metodica risulta cardinale per i seguenti risvolti clinici e socio-economici: prevenzione delle malattie epatiche evolutive, quali epatocarcinoma e cirrosi epatica; l'individuazione dello schema terapeutico con Peg-Interferone e Ribavirina più adeguato per ottenere la massima efficacia ed i minimi effetti collaterali; applicazione di una precoce sospensione della terapia se non efficace, limitando i costi terapeutici (25,000 €/*anno* per risposta sostenuta) non utili ai fini del conseguimento della guarigione (eradicazione virale). La fase di validazione ne potrebbe ammettere l'utilizzo in tutti i centri specialistici di trattamento dell'epatite C, uniformando lo schema di trattamento in tutto il territorio Veneto e Nazionale.

Il progetto interdisciplinare è stato realizzato grazie alla collaborazione del *Dipartimento di Ingegneria dell'Informazione*, in particolare del gruppo di ricerca dei Professori *Andrea Pietracaprina* e *Geppino Pucci*, e la *Clinica Medica 5* dell'Azienda Ospedaliera-Universitaria di Padova, a cui afferiscono le Dottoresse *Luisa Cavalletto* e *Liliana Chemello*.

Di seguito viene presentata una rassegna dei capitoli della tesi e il loro contenuto.

Capitolo 1: vengono introdotte le motivazioni per cui è necessario sviluppare un metodo computazionale che fornisca dei modelli ai medici epatologi per la predizione degli esiti della terapia per l'epatite C. Vengono inoltre definiti i tipi di risposte virologiche utilizzati nella nomenclatura standard di cui verrà fatto largo uso in questa tesi. In fine, si presenta qual'è lo stato dell'arte delle tecniche di predizione per la terapia in esame;

Capitolo 2: si discute di tipi di dati ponendo evidenza sulle due principali categorie. Inoltre, vengono spiegati i test statistici utilizzati per confrontare campioni indipendenti. Tali test verranno utilizzati per confrontare il gruppo dei pazienti che ottengono la risposta sostenuta e di quelli che non la ottengono;

Capitolo 3: si parla del problema della classificazione in generale. Vengono discusse le metriche di prestazione classiche delle tecniche di classificazione e delle metriche più specifiche per i problemi con classi sbilanciate. Si discute successivamente della cross-validation, un buon metodo per ottenere misure di prestazione effettive;

Capitolo 4: vengono proposti dei metodi per combinare le tecniche di classificazione standard in modo da ottenere accuratezze maggiori. Viene spiegato nel dettaglio il Bagging e un metodo per combinare i più modelli da lui prodotti. Ci si riferisce a tale metodo con Combine Multiple Model;

Capitolo 5: si introducono gli alberi decisionali come tecnica di classificazione. Tale tecnica è stata utilizzata come primo approccio per la soluzione del problema. Viene poi spiegato nel dettaglio come lavora l'algoritmo di induzione di alberi decisionali, perché alla base della soluzione proposta in questo lavoro;

Capitolo 6: si fornisce un approccio alternativo alla classificazione, il cui obiettivo non è predire se un'istanza di un dataset appartiene o meno ad una data classe, ma con che probabilità vi appartiene. Nel particolare vengono introdotti i Probability Estimation Tree (PET) e le tecniche per la loro validazione;

Capitolo 7: viene presentato il problema di predizione dell'esito della terapia per l'epatite C nel dettaglio. Viene fatta successivamente un'analisi preliminare della popolazione di pazienti in studio. Infine, viene ripercorso l'iter logico che ha portato ad ottenere i modelli predittivi definitivi, denominati PET incrementali;

Capitolo 8: si ripropone il percorso di pensiero fatto nel Capitolo 7 formalizzandolo matematicamente. Viene inoltre spiegato l'algoritmo di induzione di PET incrementali utilizzato per la costruzione dei modelli predittivi;

Capitolo 9: viene messa alla prova l'implementazione Java dell'algoritmo per la costruzione di PET incrementali. Si confrontano varie parametrizzazioni che ottengono modelli diversi in termini di complessità e di misure di prestazione. Vengono infine presentati e validati dei modelli predittivi selezionati dai medici epatologi per il loro reale impiego nella pratica clinica;

Capitolo 10: si conclude la tesi con i molteplici sviluppi possibili di questo lavoro su cui si dovrà concentrare l'attività di ricerca.

Si consigliano due percorsi di lettura, resi il più possibile indipendenti tra di loro: uno rivolto a coloro interessati alla soluzione adottata specificatamente per il problema medico in esame; l'altro rivolto invece a chi interessato anche al dettaglio tecnico dell'approccio utilizzato. Si consiglia la lettura dei Capitoli 1, 7 e 9 ai primi e 1, 7, 8 e 9 ai secondi. I Capitoli dal 2 al 6 possono essere utilizzati per supporto teorico alle analisi fatte negli altri capitoli.

Indice

Elenco delle figure

Elenco delle tabelle

Elenco degli algoritmi

Capitolo 1

Introduzione

L'*epatite* è un processo infiammatorio del fegato molto spesso causato da un virus. I più comuni virus epatotropi sono cinque e ci si riferisce a questi con la dicitura di tipo A, B, C, D ed E. La malattia è tanto diffusa che è stato istituito un giorno di sensibilizzazione mondiale, il *19 Maggio*. Lo slogan dell'anno corrente è stato *"Am I number 12?"* per mettere in luce che una persona su dodici al mondo è infetta dal virus dell'epatite di tipo B o C. Si stima che circa 180 milioni di persone al mondo siano infette in particolare dal virus dell'*epatite C* (*Hepatitis C Virus*, HCV) [1]. In Italia negli ultimi anni è persino stato rilevato un aumento dei casi di HCV a discapito di quelli di HBV (*Hepatitis B Virus*) [2], che con la disponibilità di un vaccino efficace dal 1991 sono notevolmente diminuiti.

Originariamente indicata come epatite *non A* e *non B* il virus dell'epatite di *tipo C* è stato identificato nel 1989 [3]. Si tratta di un flavivirus il cui meccanismo di azione richiede la sintesi di DNA grazie alla trascrizione inversa dell'RNA virale. Il decorso della malattia può essere asintomatico e in alcuni casi essa può rimanere silente per molto tempo (decenni). Persino la cronicizzazione della stessa può progredire per anni in modo lento e subdolo senza che il paziente se ne accorga [4]. È proprio l'epatite cronica una delle maggiori cause di cirrosi epatica e epatocarcinoma [1, 4]. La diffusione del virus è dovuto all'uso non appropriato di prodotti derivati dal sangue: attualmente le principali modalità di trasmissione dell'epatite C sono la tossicodipendenza, i trattamenti estetici, il piercing e il sesso non protetto. Non è da sottovalutare l'incidenza di coloro che hanno contratto la malattia in ambiente ospedaliero prima della disponibilità di un test in commercio che rilevasse efficacemente la presenza del virus [5].

La terapia contro HCV, dal 1996, si è basata prima, nell'uso di monoterapia con interferone (IFN), poi nella terapia di combinazione con ribavirina (RBV). L'interferone è un potente immunomodulatore, mentre la ribavirina è un analogo guanosinico capace di mutare la struttura proteica del virus HCV [6]. Più recentemente sono divenuti disponibili i risultati di ampi trial clinici internazionali che hanno valutato l'impiego di nuovi tipi di interferone a lunga emivita. Tale farmaco, l'*IFN-pegilato* (Peg-IFN) ha permesso la somministrazione su base settimanale grazie alla sua più lunga biodisponibilità. Nello specifico, in commercio sono disponibili due tipi di interferone-pegilato: l'interferone-pegilato alfa-2a (Peg-IFN alfa-2a) e alfa-2b (Peg-IFN alfa-2b). Essi differiscono nella struttura tridimen-

sionale e tale differenza ne modifica la farmaco-cinetica e la dinamica. La forma lineare del Peg-IFN alfa-2b permette la somministrazione di una dose pro chilo che gli standard internazionali hanno fissato a 1.5 $\mu g/kg$ a settimana, invece il trattamento ordinario con l'interferone-pegilato alfa-2a prevede una dose fissa, e non in base al peso del paziente, di 180 μg a settimana. La terapia prevede anche l'assunzione di ribavirina con una dose giornaliera di 15 mg/kg.

Esistono almeno 6 genotipi principali di HCV che vengono indicati con un numero progressivo da 1 a 6. La determinazione del genotipo di HCV ha rilevanti implicazioni cliniche perché conoscere il genotipo permette di stabilire a priori, seppur con ampio margine di variabilità, la probabilità di risposta alla terapia. Si deve far notare che la presenza di casi infetti dal virus dell'epatite C con genotipo 5 o 6 è molto rara in Italia. Inoltre, la determinazione del genotipo influenza anche la durata della terapia. Infatti il genotipo 1 e 4, notoriamente più resistenti alla terapia, hanno un trattamento di 12 mesi mentre il genotipo 2 e 3 è trattato per 6 mesi.

Durante gli ultimi anni si è cercato di studiare quali siano i fattori che permettono di ottenere una *risposta sostenuta* alla terapia antivirale. Si definisce *risposta sostenuta* o *risposta a lungo termine* l'eradicazione totale del virus alla fine del trattamento e in più il mantenimento di tale stato per sempre. Si dice che un paziente ha ottenuto la risposta sostenuta alla terapia per l'epatite C se a sei mesi dalla fine della terapia il virus non è rilevabile nell'organismo, e il paziente risulta quindi HCV-RNA negativo. Tali pazienti si definiscono *Long Term Responder*.

Come discusso, la malattia può essere asintomatica: solo il 30% dei casi presentano dei sintomi nella fase acuta [7], cioè nei primi 6 mesi da quando si è stati contagiati dal virus. I principali sintomi sono malessere, perdita dell'appetito e ittero ma gli effetti più devastanti sono quelli causati dalla cronicizzazione della malattia. Solo 15% dei casi si risolve spontaneamente senza arrivare alla fase cronica mentre il restante 85% dei pazienti rimane infetto ed incorre in una malattia epatica evolutiva la quale ha alti costi di gestione clinica e alte percentuali di mortalità. Circa il 20% dei pazienti cronici sviluppa la cirrosi epatica: il tessuto sano del fegato viene sostituito in buona parte con tessuto fibrotico o necrotico compromettendo le funzioni dell'organo. Tale patologia, specialmente la cirrosi indotta da HCV, è uno dei principali motivi di trapianto di fegato in Italia [8]. Per ultimo, lo stadio più progredito della malattia può portare alla generazione di tumore primitivo del fegato (*Hepatocellular Carcinoma*, HCC) di cui la cirrosi è responsabile dell' 80% circa dei casi [9].

È doveroso puntualizzare che la terapia per l'epatite C causa nel paziente molteplici effetti collaterali provocando uno stato di generale sofferenza. In molti casi, una persona affetta da questa malattia da anni, pur non presentando particolari sintomi, è costretta ad intraprendere una terapia che ne deteriora la qualità di vita. Il principale aspetto negativo dell'assunzione di interferone alfa è infatti la permanenza dalle 6 alle 8 ore dall'iniezione di uno stato simil-influenzale. Sebbene questo sintomo si attenui dopo 1 o 2 settimane di trattamento, comunemente si acutizzano il malessere, l'apatia, la mialgia, la tachicardia e l'anoressia [7]. Sempre legata a questo pesante trattamento è la comparsa di comportamenti depressivi nel paziente.

Risulta perciò importante sviluppare un *metodo computazionale* semplice ed efficace per predire la risposta sostenuta dell'individuo alla terapia. La costruzione di una terapia mirata rispetto alle varie tipologie di pazienti è il principale obiettivo di tale metodo, in modo da ottenere la massima efficacia della terapia e la riduzione al minimo degli effetti collaterali. Inoltre, identificare i pazienti ai quali possa essere sospeso il trattamento precocemente non sarebbe utile solo per evitare al soggetto inutili effetti collaterali ma permetterebbe anche di ridurre i costi della terapia somministrata inutilmente in quanto non efficace. Si fa notare che un ciclo di terapia antivirale con Peg-IFN e RBV della durata di 12 mesi costa circa 25,000 € per ogni risposta sostenuta conseguita. Nonostante gli alti costi legati alla terapia si vuole sottolineare che la spesa annua per singolo paziente non trattato in stadio non avanzato di epatopatia è di 350-450 € l'anno, che cresce nel caso di complicanze o cirrosi a 750-5000 € l'anno. Un singolo trapianto di fegato costa indicativamente dai 100,000 € ai 200,000 € [6] con una numerosità di 800 casi all'anno in Italia. Se queste spese vengono moltiplicate per il bacino di pazienti trattati per l'HCV si nota quanto questa patologia gravi sul sistema sanitario nazionale. Risulta quindi necessario ottimizzare al meglio la terapia per ridurre tali costi.

I principali indicatori dell'efficacia del trattamento si basano sul comportamento della cinetica della *viremia* (o *carica virale*) durante la terapia. I test per rilevare la viremia di un soggetto si basano su tecniche di biologia molecolare focalizzate sull'identificazione degli acidi nucleici, quale ad esempio la *Polymerase Chain Reaction* (PCR). L'unità di misura standard della carica virale sono le *Unità internazionali* su millilitro.

Per avere un riferimento alla nomenclatura standard dei tipi di risposta virologica conseguiti dal paziente vengono riportate in Tabella 1.1 le linee guida definite dall'*American Association for the Study of Liver Disease* nel 2009 [1] (AASLD). Nelle linee guida dell'AASLD è possibile estrapolare anche quali sono i parametri predittivi della terapia attualmente osservati dai medici. Il fattore più importate in assoluto si dimostra il genotipo di HCV. I genotipi 2 e 3 sono forme più sensibili che possono ottenere alte percentuali di risposta sostenuta (superiore all'80%), rispetto al genotipo 1 e 4 che sono più casi più difficili da guarire (40%). In secondo luogo risulta importante la viremia o carica virale che è espressione dell'attività replicativa del virus e se superiore a 600,000 UI/mL si dimostra fattore sfavorevole all'efficacia della terapia. Altre caratteristiche favorevolmente associate all'eradicazione virale sono: il sesso femminile, l'età più giovane (inferiore ai 40 anni), il peso inferiore a 75 kg, il livello alto di ALT (Alanina Aminotransferasi, un enzima della classe delle transaminasi contenuto all'interno delle cellule) e l'assenza di cirrosi o fibrosi come pure di insulino-resistenza.

Come discusso, la probabilità di risposta alla terapia antivirale con Peg-IFN e RBV nell'epatite cronica C dipende però, non solo dai fattori sopracitati, ma è soprattutto determinata dal comportamento della cinetica virale durante la terapia. Per esempio, non raggiungere la soppressione di 2 logaritmi al terzo mese rispetto ai valori basali, cioè non ottenere l'EVR, comporta nella totalità dei casi il non riuscire ad ottenere la LTR. Per contro, ottenere l'EVR non predice necessariamente la risposta sostenuta e in questo caso si dovrà tener conto del completamento della corretta posologia terapeutica (dose standard di trattamento). Infine, ottenere una rapida risposta alla terapia, cioè già al primo mese

Tipo di risposta virologica	Definizione
Rapid Virological Response (**RVR**)	HCV-RNA negativo al primo mese di trattamento. La quantità di virus, misurabile in UI/mL tramite uno strumento basato sulla PCR (*Polymerase Chain Reaction*), deve essere al di sotto della soglia di rilevabilità
Early Virological Response (**EVR**)	$\log_{10}(\textit{viremia basale}) - \log_{10}(\textit{viremia } 3° \textit{mese}) \geq 2$ Decremento di almeno due logaritmi della vireamia al terzo mese rispetto alla viremia basale
complete Early Virological Response (**cEVR**)	Completa eradicazione virale al terzo mese di terapia
partial Early Virological Response (**pEVR**)	Completa negativizzazione della viremia al quarto, quinto o sesto mese
End-of-Treatment Response (**ETR**), *Response* (**R**)	Soppressione del virus a fine terapia. Viene sempre valutata la viremia alla fine della terapia sia che questa duri 6 mesi sia che duri 12 mesi o più
Sustained Virological Response (**SVR**), *Long-Term Response* (**LTR**)	Virus non rilevabile nemmeno dopo 6 mesi dalla fine del trattamento
Relapse (**RR**)	Il virus era stato soppresso con la terapia ma dopo 6 mesi dalla stessa si ripresenta
Not Response (**NR**)	Virus sopra la soglia di rilevabilità al sesto mese di trattamento

Tabella 1.1: Definizioni dei tipi di risposte virologiche

(RVR) si dimostra il più forte predittore della LTR per tutti i genotipi, anche per quello più resistente di tipo 1.

In questa tesi viene fornito un approccio ragionato che tiene conto di tutti parametri predittivi e di come sono legati tra di loro per predire la risposta sostenuta alla terapia ad un qualsiasi istante temporale di trattamento. L'obiettivo finale è la costruzione di tre *modelli* relativi rispettivamente allo stato basale (prima del trattamento) del paziente, al primo e al terzo mese di terapia. Semplicemente osservando i parametri che meglio predicono l'esito della terapia lo specialista sarà in grado di stimare la probabilità di risposta sostenuta al trattamento dell'epatite C per un nuovo paziente. I parametri selezionati per la costruzione del modello, dal metodo computazionale che lo costruisce, sono misurabili tramite semplici test clinici di routine. Questa è la carattestica che li rende realmente utilizzabili in pratica. La natura altamente dinamica del problema ha permesso di spingersi oltre la sola predizione. I modelli costruiti permettono di ottenere la probabilità di guarigione in base alle informazioni note del nuovo paziente in un dato istante temporale e di osservare come essa possa variare nel futuro a seconda delle *scelte* che il medico potrebbe compiere. Modelli di questo tipo sono di utile supporto alle decisioni critiche che vengono prese al pretrattamento, al primo e al terzo mese. Il medico avrà la possibilità di basarsi su informazioni specifiche per scegliere se è il caso di indurre uno stop prematuro alla terapia a causa della sua poca efficacia. Inoltre, potrà valutare facilmente che risultati potrebbe apportare un eventuale scalo della dose di farmaco dovuto ad effetti collaterali. Dal pun-

to di vista farmaco-economico questo metodo può migliorare il rapporto di costo-efficacia della terapia per l'HCV.

Nello specifico i modelli ottenuti sono *alberi decisionali per la stima di probabilità* costruiti con un approccio *incrementale* per serapare logicamente il passato e il futuro della terapia ad un dato istante temporale. Allo stato dell'arte non sono presenti in letteratura modelli predittivi chiari e semplici della cinetica virale dell'HCV. Modelli dinamici della cinetica virale sono stati proposti in [10, 11] ma l'approccio troppo teorico ne rende difficile l'applicazione alla pratica comune. La stima della probabilità di risposta sostenuta è stata invece presa in considerazione in [12]. Vengono proposte delle formule da utilizzare ad istanti temporali diversi della terapia per stimare la probabilità di riuscita della stessa. Tali formule vengono costruite con la regressione logistica multivariata. Il risultato è interessante, viste le accuratezze dei modelli, ma rimane ancora poco trasparente e di poco interesse per la pratica clinica. In [13] vengono invece proposti alberi decisionali per stimare la probabilità di ottenere la RVR o la cEVR allo stato basale del paziente, cioè prima dell'inzio della terapia. Purtroppo l'informazione fornita dal modello è solo parziale in quanto viene stimata la probabilità di ottenere un fattore predittivo della risposta sostenuta (quale la RVR o la cEVR) e non la probabilità di ottenere la risposta sostenuta vera e propria. Osservando anche questi lavori, si è sviluppato un metodo ad hoc per la costruizione di modelli che avessero buone prestazione e che fossero semplici e chiari da comprendere.

Il risultato è il frutto della collaborazione tra il *Dipartimento di Ingegneria dell'Informazione*, in particolare del gruppo di ricerca dei Professori *Andrea Pietracaprina* e *Geppino Pucci*, e la *Clinica Medica 5* dell'Azienda Ospedaliera-Universitaria di Padova a cui afferiscono le Dottoresse *Luisa Cavalletto* e *Liliana Chemello*.

Gli alberi decisionali per la stima di probabilità ottenuti con l'approccio incrementale sono stati valutati con le ROC curve. I valori *minimi* delle aree sottese dalle curve ROC (AUC) nei test effettuati nel Capitolo 9 sono stati 0.593, 0.753 e 0.819 rispettivamente allo stato basale del paziente (pretrattamento), al primo mese e al terzo mese di terapia. I valori *massimi* nello stesso ordine sono stati 0.668, 0.779 e 0.895. Le AUC sono state calcolate mediando 10 AUC ognuna ottenuta con una 10-fold cross-validation. Questi valori certificano la validità dei modelli proposti.

Il risultato di interesse medico è presentato in maniera totalmente indipendente dalla formalizzazione matematica del problema. Pertanto, il percorso di lettura consigliato a coloro interessati ai soli modelli predittivi proposti è il Capitolo 7 e 9 che presentano rispettivamente l'analisi del problema clinico e i risultati sperimentali. Per chi è interessato invece all'aspetto tecnico si consiglia la lettura dei Capitoli 7, 8 e 9. Il Capitolo 8 presenta appunto la formalizzazione matematica del problema. I Capitoli dal 2 al 6 possono essere utilizzati per supporto teorico ai capitoli con le analisi dei risultati.

Capitolo 2

Test statistici

In questo capitolo si vuole per primo definire formalmente cos'è un attributo, che è un'informazione base per comprendere la trattazione fatta in questo lavoro di tesi. Questo viene fatto nel Paragrafo 2.1. Una volta definiti i due tipi principali di attributi, cioè quelli *categorici* e quelli *numerici*, si procede proponendo una breve teoria sui test statistici relativi ad essi nel Paragrafo 2.2. Viene posta particolare attenzione ai test che permettono di valutare come si comporta un attributo in due popolazioni distinte.

2.1 Tipi di dato

Un *dataset* può essere visto come un insiemi di oggetti descritti da un insieme di *attributi* che ne catturano le caratteristiche base. Un attributo può essere comunemente chiamato anche *parametro* o *variabile*.

Definizione 2.1. *Un attributo è una proprietà o una caratteristica di un oggetto che può variare.*

Ogni attributo viene categorizzato differentemente a seconda del suo *tipo*. Il tipo di un attributo è utile per identificare le sue proprietà e i modi adeguati con cui maneggiarlo. Per riconoscere il tipo di un attributo si devono specificare le proprietà dei valori assunti da esso, infatti a seconda delle possibili operazioni che si possono applicare a tali valori esso viene categorizzato in modo diverso. In Tabella 2.1 sono elencate le varie tipologie di attributi. È necessario sottolineare che a volte basta discriminare i parametri *categorici* da quelli *numerici* senza dover andare al livello superiore di dettaglio evidenziato nella tabella appena introdotta.

2.2 Test di ipotesi

Il valore di un attributo può essere altamente variabile in una popolazione: è utile perciò studiare la probabilità con cui un attributo assume i suoi valori. La branca della statista che si pone domande sulla *distribuzione di probabilità* di una variabile nella popolazione a

Attributi Categorici/Nominali	
Nominali:	il cui dominio, di solito discreto, comprende valori sui quali *non* è definibile un ordinamento significativo. *Insieme delle operazioni ammissibili*: $\{=, \neq\}$ *Esempi*: codice fiscale, colori.
Ordinali:	il cui dominio, di solito discreto, comprende valori sui quali è definito un ordinamento totale e significativo. *Insieme delle operazioni ammissibili*: $\{=, \neq <, \leq, >, \geq\}$ *Esempi*: giudizi, numeri civici.
Attributi Numerici/Quantitativi	
Discreti:	il cui dominio è numerico e discreto. *Insieme delle operazioni ammissibili*: $\{=, \neq <, \leq, >, \geq +, -\}$ *Esempi*: età.
Continui:	il cui dominio è numerico e continuo. *Insieme delle operazioni ammissibili*: $\{=, \neq <, \leq, >, \geq +, -, \times, :\}$ *Esempi*: temperatura, peso.

Tabella 2.1: Tipologie di attributi

partire da un campione di cui si è a disposizione è la *Statistica Inferenziale*. Si pensi di avere un campione di 100 persone con una determinata malattia di cui ne guariscono solo 20. È possibile concludere che la probabilità di guarigione sia il 20%? La risposta a domande di questo tipo viene data in libri di statistica classica, quale ad esempio [14], non si vuole quindi andare in ulteriore dettaglio. Si vuole invece parlare di un campo particolare della statistica inferenziale: i *test di ipotesi*. Più precisamente i test di ipotesi per il confronto di due campioni indipendenti. Test del genere vengono utilizzati spesso in ambito medico, infatti in questo lavoro vengono utilizzati per confrontare gli attributi tra il campione dei pazienti che rispondono alla terapia e gli altri. I test di ipotesi che vengono introdotti in questo paragrafo sono specifici e per una trattazione estesa è consigliabile leggere [15]. Un test di ipotesi si sviluppa in 6 passi:

1. Analisi dei dati disponibili;
2. Verifica degli assunti di base;
3. Impostazione dell'ipotesi statistica (definita *ipotesi nulla* H_0);
4. Costruzione della statistica;
5. Determinazione della distribuzione statistica;
6. Definizione della regola di decisione.

Ogni test lavora su un certo tipo di dati e fa delle assunzioni di base sulla loro distribuzione, è pertanto necessario identificare il test giusto per un determinato parametro. Dopodiché viene formulata l'ipotesi che si vuole testare, la cosiddetta ipotesi nulla (H_0), e costruita la statistica per prendere le decisioni riguardo all'ipotesi formulata. La *statistica* è semplicemente una variabile aleatoria in funzione del del campione in esame. Se l'ipotesi è corretta tale variabile aleatoria ha una certa distribuzione, è proprio in base ad essa che si accetta o si rifiuta l'ipotesi H_0. In particolare l'ipotesi nulla viene rifiutata se si è al di sotto di una determinata probabilità d'errore nel farlo stabilita a priori: tale soglia viene definita *livello di significatività*. La probabilità di commettere un errore nel rifiutare l'ipotesi nulla è un valore cruciale e viene denominato *p-value*. Solitamente l'ipotesi nulla fa l'assunzione che non vi sia differenza tra i due campioni. Pertanto un valore di p basso indica che l'attributo in esame ha distribuzione significativamente diversa nei due campioni. Per questo motivo, spesso nelle tabelle relative ai test statistici di confronto vengono presentate delle informazioni relative ad un attributo nei due campioni in esame e il p-value del test associato ad esso.

2.2.1 Test del Chi-quadro

Il test del *Chi-quadro* è stato introdotto da Karl Pearson e per questo la sua versione più semplice viene denominata *Chi-quadro di Pearson*. Questo test di ipotesi può essere utilizzato per il confronto di frequenze di risposte binarie in due campioni indipendenti. In tali casi è utile costruire una tabella a doppia entrata chiamata *tabella di contingenza*. Per ognuno dei due gruppi deve essere riportato il conteggio delle risposte binarie. Il test del Chi-quadro permette di verificare scientificamente se le proporzioni tra i due valori nei due gruppi sono dovute al caso o meno.

Si supponga di voler studiare le frequenze dei valori dell'attributo binario A, il cui dominio è $\{x, y\}$, in due campioni C_1 e C_2 indipendenti. Con il riferimento alla tabella di contingenza in 2.2 la domanda che ci si pone è se la differenza tra $\frac{a}{c}$ e $\frac{b}{d}$ sia dovuta al caso o è significativa. L'ipotesi nulla (H_0) stabilita dal test del Chi-quadro è che la differenza

	C_1	C_2	
elementi con $A = x$	a	b	$\mathbf{n_1}$
elementi con $A = y$	c	d	$\mathbf{n_2}$
	$\mathbf{n_3}$	$\mathbf{n_4}$	$\mathbf{N}$

Tabella 2.2: Tabella di contingenza

sia dovuta esclusivamente a fattori casuali, l'ipotesi alternativa (H_1) è che la differenza sia reale anche se le sue cause sono ignote. Nel caso non vi fosse nessuna differenza per l'attributo A nei due gruppi i rapporti $\frac{a}{c}$ per C_1 e $\frac{b}{d}$ per C_2 dovrebbero essere uguali. Viene così costruita un tabella di contingenza che mantiene uguali tali frequenze tenendo conto della diversa numerosità dei campioni (Tabella 2.3). La statistica viene poi così costruita:

$$x = \frac{(a - a')^2}{a'} + \frac{(b - b')^2}{b'} + \frac{(c - c')^2}{c'} + \frac{(d - d')^2}{d'}$$

	C_1	C_2	
elementi con $A = x$	$a' = n_1 \cdot n_3/N$	$b' = n_1 \cdot n_4/N$	$\mathbf{n_1}$
elementi con $A = y$	$c' = n_2 \cdot n_3/N$	$d' = n_2 \cdot n_4/N$	$\mathbf{n_2}$
	$\mathbf{n_3}$	$\mathbf{n_4}$	**N**

Tabella 2.3: Tabella di contingenza attesa nel caso in cui le differenze nelle frequenze siano casuali

Se l'ipotesi nulla è corretta, cioè le differenze sono dovute al caso, la statistica ha una distribuzione di tipo $\chi^2_{(1)}$. La probabilità di sbagliare rifiutando l'ipotesi H_0 è data da $1 - F(x)$ dove $F(\cdot)$ è la funzione di ripartizione di una variabile $\chi^2_{(1)}$. Tale probabilità è detta p-value e se è al di sotto del livello di significatività scelto viene rifiutata l'ipotesi nulla. Per rendere chiaro il procedimento viene proposto l'Esempio 2.1.

Esempio 2.1. Si vuole studiare se l'attributo binario "Sesso" ha una distribuzione di probabilità uguale nel gruppo dei pazienti che rispondono ad una terapia e nel gruppo di quelli che non rispondono. Viene allora generata la tabella di contingenza relativa all'attributo nei due gruppi e la tabella di contingenza attesa nel caso le differenze delle distribuzioni di probabilità dell'attributo siano casuali, cioè nel caso valga H_0.

	Responsivi	Non responsivi	
Femmina	126	75	**201**
Maschio	237	168	**405**
	363	**243**	**606**

	Responsivi	Non responsivi	
Femmina	120	81	**201**
Maschio	243	162	**405**
	363	**243**	**606**

Viene calcolata la statistica:

$$x = \frac{(126-120)^2}{120} + \frac{(75-81)^2}{81} + \frac{(237-243)^2}{243} + \frac{(168-162)^2}{162} \approx 1.115$$

A questo punto la probabilità di sbagliare nel dire che la percentuale di donne presenti nei responsivi di 0.347 e quella presente nei non responsivi di 0.309 sia dovuta al caso è $1 - F(x) \approx 0.324$. Scelta la soglia di significatività $\alpha = 0.05$ l'ipotesi nulla H_0 non viene rifiutata. ▪

2.2.2 Test di Student

Il test di *Student* viene utilizzato per confrontare il valore medio di un attributo numerico in due campioni indipendenti. Gli attributi numerici possono essere considerati infatti come variabili aleatorie a media μ e varianza σ^2. A differenza del test del Chi-quadro il test di Student fa delle assunzioni a priori sulla distribuzione di probabilità dei campioni in esame:

deve essere di tipo *gaussiano*. Tale ipotesi è verificata nella maggior parte dei casi ma non sempre. Inoltre si vuole che le varianze siano uguali in entrambi i gruppi. L'ipotesi nulla da sottoporre a verifica è quindi $H_0 : \mu_{C_1} = \mu_{C_2}$, dove μ_{C_1} è il valor medio dell'attributo A in esame nella popolazione relativa al campione C_1 e μ_{C_2} è il valor medio della popolazione relativa al campione C_2. Naturalmente l'ipotesi alternativa è $H_1 : \mu_{C_1} \neq \mu_{C_2}$. La statistica viene così costruita:

$$t = \frac{m_{C_1} - m_{C_2}}{s_p \cdot \sqrt{\frac{1}{N_{C_1}} + \frac{1}{N_{C_2}}}}$$

I parametri m e N sono rispettivamente la media del campione e la sua numerosità, il pedice invece indica il campione a cui ci si riferisce. Il valore di s_p è il seguente

$$s_p = \sqrt{\frac{(N_{C_1} - 1)s_{C_1}^2 + (N_{C_2} - 1)s_{C_2}^2}{N_{C_1} + N_{C_2} - 2}}$$

che è semplicemente una media ponderale delle varianze $s_{C_1}^2$ e $s_{C_2}^2$ dei due campioni. Se vale l'ipotesi nulla la statistica è una variabile aleatoria con distribuzione nota, detta di Student, la cui distribuzione dipende unicamente dai gradi di libertà $d = N_{C_1} + N_{C_2} - 2$. È possibile calcolare il p-value conoscendo la funzione di ripartizione $F(\cdot)$ per una variabile aleatoria di Student a d gradi di libertà così $p = 2(1 - F(t))$. Il modo di calcolare il valore di p appena introdotto è detto *two-sided* perché tiene conto di entrambe le code della distribuzione di probabilità della statistica. Se tale valore è al di sotto della soglia di significatività α scelta viene rifiutata l'ipotesi nulla H_0 e accettata implicitamente H_1. Si sceglie di presentare l'esempio 2.2 per completezza.

Esempio 2.2. Si vuole studiare la distibuzione di probabilità dell'attributo continuo "Età" (in anni) nel campione di pazienti responsivi ad una malattia (C_1) e in quello dei pazienti non responsivi (C_2). Rispettivamente il primo è composto da 157 persone (N_{C_1}) ed il secondo da 195 (N_{C_2}). Le medie e le deviazioni standard nei due gruppi sono le seguenti:

$$m_{C_1} = 44,\ m_{C_2} = 48,\ s_{C_1} = 10,\ s_{C_2} = 11$$

Ora che si hanno tutti i dati è possibile calcolare il valore di t come spiegato nella parte teorica con

$$t = \frac{44 - 48}{s_p \cdot \sqrt{\frac{1}{157} + \frac{1}{195}}}$$

avendo prima calcolato s_p con la formula

$$s_p = \sqrt{\frac{(157 - 1)10^2 + (195 - 1)11^2}{157 + 195 - 2}}$$

Si ottiene $t \approx -3.604$ che permette di calcolare il valore di $p = 2(1 - F(t)) \approx 0.000$ approssimato a 3 cifre significative. Scegliendo un livello di significatività di 0.05 si deve rifiutare l'ipotesi nulla, che specificava che la popolazione di C_1 ha lo stesso valor medio μ di quella di C_2. ▪

Fissato un livello α di significatività i valori di p più piccoli di tale soglia non vengono scritti ma viene indicato $p \leq \alpha$ per indicare che si è rifiutato l'ipotesi nulla commettendo un errore che è al massimo α.

Capitolo 3

Classificazione

Tutte le formalizzazioni matematiche più complesse di questa tesi utilizzano la notazione introdotta in questo capitolo. Nel Paragrafo 3.1 viene spiegato rigorosamente cosa si intende per classificazione. Il resto del capitolo è incentrato sulle metriche per valutare la qualità delle tecniche di classificazione: le metriche classiche vengono proposte nel Paragrafo 3.2 e quelle specifiche per problemi a classi sbilanciate nel Paragrafo 3.4. Si ponga particolare attenzione al Paragrafo 3.3 visto che è la tecnica che verrà usata per validare i modelli prodotti nel lavoro di tesi.

3.1 Definizione del problema

Di seguito vengono introdotte alcune definizioni con lo scopo di utilizzare una notazione standard in tutto il lavoro. È comodo rappresentare i dati con il *modello relazionale* la cui definizione formale è basata sulla teoria matematica degli insiemi. Per maggiori informazioni su tale modello si consulti [16].

Un *training set* T è una relazione composta da n *tuple* (che possono essere chiamate anche *record*, *instance* o *example*) ognuna delle quali è costituita da un elenco ordinato di m attributi ed 1 classe. Ogni attributo A_j valori sul rispettivo dominio D_j con $1 \leq j \leq m$. Sempre per comodità di notazione ci si riferirà all'insieme di tutti gli attributi (che possono essere chiamati anche *feature*) con la lettera corsiva $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$. La classe C prende invece valori su $\Gamma = \{\gamma_0, \gamma_1, \ldots, \gamma_{c-1}\}$ che rappresenta l'insieme delle c classi disponibili, con c numero finito. Comunemente ci si riferisce alla classe C anche con il termine *category* o *target attribute*.

È ora chiaro che, sempre in riferimento al modello relazionale, $T \subseteq D_1 \times D_2 \times \cdots \times D_m \times \Gamma$ sullo schema di attributi $\mathcal{A} \cup C$ dove C è il target attribute. Più in generale invece un *dataset* D può essere costituito da record la cui classe non è nota, quindi $D \subseteq D_1 \times D_2 \times \cdots \times D_m$ sullo schema di attributi $\mathcal{A}$. La relazione può essere rappresentata graficamente in forma di tabella, la rappresentazione tabellare di una relazione è molto intuitiva e facilmente comprensibile. La tabella per il training set T appena introdotto è costituita da n righe e $m + 1$ colonne ed è raffigurata in Tabella 3.1.

A_1	A_2	...	A_m	C
$t_1[A_1]$	$t_1[A_2]$	...	$t_1[A_m]$	$t_1[C]$
$\vdots$				$\vdots$
$t_n[A_1]$	$t_n[A_2]$	...	$t_n[A_m]$	$t_n[C]$

Tabella 3.1: Tabella che rappresenta il training set T

Ora che si sono introdotti i concetti fondamentali è possibile definire il problema di classificazione in maniera formale.

Problema 3.1 (Classificazione)**.**
INPUT: *Training set T, dove $T = \{t_1, t_2, \ldots, t_n\}$ è l'insieme di n record.*

OUTPUT: *Target function o Classification model M che mappi ogni tupla del prodotto cartesiano $D_1 \times D_2 \times \cdots \times D_m$ in Γ.*

$$M : D_1 \times \cdots \times D_m \longrightarrow \Gamma$$
$$t \longmapsto M(t)$$

L'approccio sistematico per risolvere il Problema 3.1 è definito *classification technique* o *classifier*. Ogni tecnica utilizza un algoritmo di learning (*learning algorithm*) per identificare il modello che meglio approssima la relazione tra gli attributi in input e la classe. In questo lavoro ci si riferirà ad un learning algorithm con $\mathcal{L}$. In una classification technique è prevista anche una fase di *validazione* che mira a valutare la qualità del modello o dei modelli trovati. Per poterla espletare è necessario un dataset con la stessa struttura del training set utilizzato nella fase di learning, tale insieme viene comunemente chiamato *test set*. Dei record appartenenti al test set $\widetilde{T}$ deve essere nota la classe per poterla confrontare con quella *predetta* dal modello. In Figura 3.1 è riassunto l'approccio generale alla classificazione.

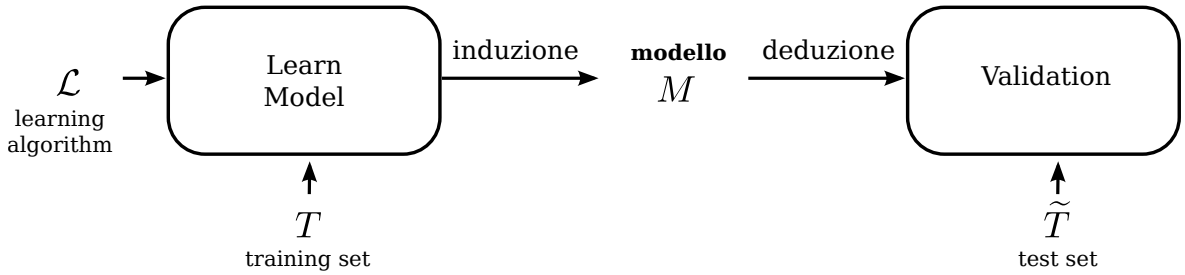


Figura 3.1: Approccio generale alla classificazione

3.2 Metriche di prestazione

Per valutare la qualità di un *classification model M* non basta verificare che classifichi correttamente i record appartenenti all'insieme di training da cui è stato costruito, ma

deve classificare accuratamente anche record mai elaborati prima. Si deve pertanto fornire una stima della bontà della classificazione del modello per record generici.

Sia U l'universo dei record che si vogliono studiare ed $E \subseteq U$ il *training set*, cioè l'insieme dei record grazie al quale è stato costruito il modello.

$$M(t) \equiv \textit{classe prevista dal modello } M \textit{ per } t$$

e

$$t[C] \equiv \textit{classe vera del record } t$$

per introdurre i seguenti tipi di errore:

- *training error*, $t \in E : \ M(t) \neq t[C]$;
- *generalization error*, $t \in U \setminus E : \ M(t) \neq t[C]$.

Il *training error rate* è facilmente calcolabile così

$$\frac{|\{t \in E : \ M(t) \neq t[C]\}|}{|E|}$$

il *generalizzation error rate* è invece solo stimabile, in quanto si vuole descrivere un fenomeno aleatorio quale la percentuale di record classificati erroneamente dal modello tra quelli non ancora osservati. Naturalmente l'*accuratezza* del modello (*accuracy*) è così definita:

$$\textit{accuracy} = 1 - \textit{generalization error rate}$$

3.3 Cross-Validation

Per ottenere una stima corretta dell'accuratezza della classification technique è possibile utilizzare varie tecniche consolidate e discusse in letteratura (si veda [17] cap. 4.5). In [18] è stato effettuato un ampio studio per mettere a confronto metriche di prestazione ottenute con metodi di stima diversi: i risultati suggeriscono che il miglior metodo di stima è una *10-fold cross-validation*.

Dato un *classification model* M e un insieme la cui classe dei record è nota (E) è bene quindi calcolare le metriche di prestazione con la *cross-validation*. Il primo passo è quello di partizionare l'insieme E in k sottoinsiemi

$$E = E_1 \cup E_2 \cup \cdots \cup E_k$$

tali che

$$E_i \cap E_j = \emptyset \ \forall \ i \neq j \ 1 \leq i, j \leq k$$

con taglia

$$|E_i| = \frac{1}{k}|E| \ \forall i : \ 1 \leq i \leq k$$

Viene quindi costruito un modello M_i sul training set $E \setminus E_i$ per poi utilizzare E_i come test set, questo procedimento viene iterato k volte. Le accuratezze ottenute dai k modelli vengono mediate per ottenere una misura di accuratezza complessiva.

Quindi l'accuratezza stimata nella cross-validation è:

$$\frac{1}{k}\sum_{i=1}^{k}\frac{|\{t \in E_i :\ M_i(t) = t[C]\}|}{|E_i|} = \frac{\sum_{i=1}^{k}|\{t \in E_i :\ M_i(t) = t[C]\}|}{|E|}$$

Il processo appena descritto prende il nome di *k-fold cross-validation*.

3.4 Ulteriori metriche di prestazione

Per quantificare la qualità della classificazione oltre alla misura dell'accuratezza è necessario introdurre metriche alternative per valutare la bontà nella classificazione di record appartenenti ad una classe specifica. Per rendersi conto dell'inadeguatezza della misura *accuracy* nel caso di classi sbilanciate vengono proposti i seguenti esempi: un nell'ambito fraud detection e l'altro nel settore medico.

Esempio 3.1. Si supponga di voler identificare con un classificatore le transazioni fraudolente in un database che contiene transazioni di carta di credito. Sperabilmente queste saranno meno dell'1% del totale. Utilizzando un classificatore che predice ogni transazione come legittima si otterrebbe un'accuratezza altissima pari al 99% sebbene fallisca nel trovare quelle fraudolente. ▪

Esempio 3.2. Volendo classificare i pazienti che rispondono alla terapia combinata di interferone (IFN) e ribavirina (RBV) è possibile commettere un errore se si osserva solo l'accuratezza: nel primo database messo a disposizione per questo lavoro 188 pazienti rispondono alla terapia ("Responder") e 55 non rispondono ("Not Responder"), l'approccio naïf di associare ogni record alla classe "Responder" permette di ottenere un'accuratezza di $\frac{188}{243} \approx 0.774$. ▪

In un problema di classificazione binaria è possibile denotare una classe come *positiva* e una come *negativa*. Di seguito viene introdotta la terminologia in riferimento alla *confusion matrix* sottostante.

		Classe predetta	
		+	-
Classe reale	+	f_{++} (TP)	f_{+-} (FN)
	-	f_{-+} (FP)	f_{--} (TN)

- True Positive, f_{++} (TP) - numero dei record positivi classificati correttamente nella classe positiva;
- False Negative, f_{+-} (FN) - numero dei record positivi classificati erroneamente nella classe negativa;

- False Positive, f_{-+} (FP) - numero dei record negativi classificati erroneamente nella classe positiva;
- True Negative, f_{--} (TN) - numero dei record negativi classificati correttamente nella classe negativa.

Viene definita la *sensibilità* o *True Positive Rate* come la frazione dei record positivi classificati correttamente come positivi

$$TPR = \frac{TP}{TP + FN}$$

la *specificità* o *True Negative Rate* come la frazione dei record negativi classificati correttamente come negativi

$$TNR = \frac{TN}{TN + FP}$$

la *False Positive Rate* come la frazione dei record negativi classificati come positivi

$$FPR = \frac{FP}{FP + TN}$$

e la *False Negative Rate* come la frazione dei record positivi classificati come negativi

$$FNR = \frac{FN}{FN + TP}$$

Altre misure utilizzate sono la *precision* che definisce la frazione di record realmente positivi tra quelli classificati come positivi dal modello

$$\textit{precision},\ p = \frac{TP}{FP + TP}$$

e la *recall* che per definizione è uguale al *true positive rate*

$$\textit{recall},\ r = \frac{TP}{TP + FN} = TPR$$

Più alta è la *precision* meno sono i falsi positivi commessi dal modello, un alto *recall* indica invece che i falsi negativi sono limitati. Ad esempio classificando ogni record come positivo si ottiene recall 1 ma precision molto bassa, altrimenti se ogni record preso dall'insieme test viene classificato come positivo solo se identico ad un record dell'insieme di training si ottiene precision 1 ma recall molto bassa. La recall misura la capacità di individuare *tutti* i record positivi mentre la precision misura la capacità di individuare *soltanto* quelli positivi. Strategie che mirano ad ottenere alta precision diminuiscono la recall e viceversa, naturalmente quello che si vorrebbe è massimizzarle entrambe.

Una misura che le racchiude entrambe è la *F-measure*:

$$F = \frac{2}{\frac{1}{r} + \frac{1}{p}} = \frac{2rp}{r + p} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

La *F-measure* rappresenta la media armonica tra precisione e richiamo, questa tende ad essere più vicina al più piccolo tra i valori.

È utile calcolare le confusion matrix anche con la cross-validation, se si utilizzasse il solo training set per calcolarla si avrebbe avrebbe una stima ottimistica della prestazione del modello. Come per la stima dell'accuratezza (nel Paragrafo 3.3) il training set E viene partizionato in k sottoinsiemi. Vengono eseguite k iterazioni ad ognuna delle quali viene costruito un modello M_i sull'insieme di record $E \setminus E_i$ e utilizzato E_i come test set per tale modello. Per ogni record del test set E_i viene identificata la classe reale e quella predetta per incrementare il valore della cella corretta nella confusion matrix. Alla fine del procedimento la somma delle celle della confusion matrix sarà uguale alla taglia di E.

Capitolo 4

Classificazione combinata o Ensemble

Nel Paragrafo 4.1 si discute di come possa essere migliorata l'accuratezza nella classificazione combinando diverse classification technique. Ci si sofferma poi nel particolare sulla spiegazione dettagliata del Bagging nel Paragrafo 4.2, perché utilizzato nel tentativo di migliorare le predizione dei modelli in questo lavoro di tesi. Tale tecnica è stata provata anche come parametro del Combine Multiple Models trattato nel Paragrafo 4.3.

4.1 Introduzione alla tecnica

Il metodo *ensemble* stabilisce l'utilizzo di un insieme di *classificatori base* che poi vengono combinati per ottenere un minor errore nella predizione. Un metodo classico per combinarli, per la classificazione di un nuovo record t, è utilizzare un sistema di voto che conti quanti classificatori base assegnerebbero t ad ognuna delle possibili classi. Naturalmente t viene assegnato alla classe che ottiene più voti. La struttura base di questa tecnica è mostrata nell'Algoritmo 4.1, l'ipotesi è quella di possedere un traning set T. Si sottolinea che nell'Algoritmo 4.1 viene solo mostrata la linea di principio dei metodi ensemble, per questo non è ben specificato cosa accade alla riga 3: dipende dal metodo particolare che si sceglie.

Algoritmo 4.1 Classificatore composito o Ensemble

1 Siano k i modelli che si vogliono generare
2 **for** $i \leftarrow 1$ **to** k **do**
3 Ottieni il modello M_i come output
di un problema di classificazione che ha ricevuto in input
un training set manipolato da T
4 Classifica un nuova istanza t come $M^*(t) = \textsc{Combine}(M_1(t), \ldots, M_k(t))$

Una comune implementazione della funzione Combine è la seguente

COMBINE($M_1(t), \ldots, M_k(t)$)
return $\arg\max_{\gamma} \sum_{i=1}^{k} \delta(M_i(t) = \gamma)$

$\delta(\pi)$ è la funzione *indicatrice* della proposizione π che da valore 1 se la proposizione è vera e 0 altrimenti.

L'idea è semplice e gli algoritmi implementati in letteratura hanno realmente ottenuto risultati migliori che utilizzando una singola tecnica di classificazione. Una presentazione generale di queste tecniche può essere reperita in [19] e uno studio sulla reale efficacia dei classificatori compositi è stato fatto in [20]. Si prenda in considerazione il seguente esempio per rendere l'idea della bontà di questa tecnica.

Esempio 4.1. Supponendo di avere 25 modelli il cui generalization error rate è $e = 0.35 \; \forall i : 1 \leq i \leq 25$, si supponga inoltre che i classificatori siano *indipendenti* cioè che i loro errori non siano correlati tra di loro. Dato un record t non ancora classificato è possibile definire 25 variabili aleatorie $X_i \sim b(e_i)$ (di Bernoulli) i.i.d. come

$$X_i = \begin{cases} 1 & \text{se } M_i(t) = t[C] \\ 0 & \text{se } M_i(t) \neq t[C] \end{cases}$$

È inoltre noto che la somma di n variabili aleatorie di Bernoulli a parametro p ed indipendenti tra di loro producano una variabile aleatoria Binomiale a parametri n e p, quindi $X_1 + \ldots X_{25} = X \sim Bin(25, e)$. In caso che si combinino i classificatori con un sistema di voto per ottenere una classificazione sbagliata di un nuovo record bisogna che almeno più della metà dei modelli classifichino il record in una classe sbagliata, cioè

$$\mathbf{P}(M^*(t) = t[C]) = \mathbf{P}(X \geq 13) = \sum_{i=13}^{25} \binom{25}{i} e^i (1-e)^{25-i} = 0.06$$

▪

L'esempio 4.1 illustra come sia necessario che i classificatori di base siano tra di loro indipendenti. È possibile inoltre dimostrare che, per mantenere l'efficacia del metodo, è necessario che siano più accurati di un classificatore che ipotizzi la classe di appartenenza di un record in maniera random uniforme. Si preferirebbe quindi che per un problema di classificazione, dove c è il numero delle classi, i modelli che si vogliono combinare abbiano generalization error rate minore di $\frac{1}{c}$. In pratica è difficile ottenere l'indipendenza tra i modelli, ma si è riusciti comunque ad ottenere buoni risultati.

I principali metodi per costruire un classificatore composito si basano tutti sulla struttura dell'Algoritmo 4.1 e implementano in maniera specifica la costruzione dei diversi modelli in riga 3:

Manipolando i record del training set : i training set utilizzati per costruire i modelli vengono costruiti come un campione casuale del training set E fornito in input al problema. Esempi di queste tecniche sono il *Bagging* [21] e il *Boosting* [22]

Manipolando l'insieme degli attributi : dall'insieme delle feature $\mathcal{A}$ in input è possibile costruire diversi training set sullo schema di attributi $\bar{\mathcal{A}} \cup C$ dove $\bar{\mathcal{A}} \subseteq \mathcal{A}$. Un sottoinsieme delle feature può essere trovato in maniera random o accuratamente selezionando gli attributi uno ad uno. Un metodo che manipola l'insieme delle feature sono le *Random Forest* proposte in [23].

Manipolando l'insieme delle classi : si trasforma il problema di classificazione in un problema di classificazione binaria partizionando in due sottoinsiemi, ed in maniera random, l'insieme Γ delle classi. Per ogni modello M_i siano Γ'_i e Γ''_i i sottoinsiemi generati, che in questo caso diventano le due classi del problema di classificazione binario. Una volta a disposizione un nuovo record da classificare t viene aggiunto iterativamente un punto a tutti le classi di Γ'_i se $M_i(t) = \Gamma'_i$ altrimenti viene aggiunto un punto alle classi di Γ''_i. Il record t verrà assegnato alla classe che otterrà il maggior punteggio.

Esempio 4.2. Si supponga che l'insieme Γ sia costituito da 4 classi: $\Gamma = \{\gamma_0, \gamma_1, \gamma_2, \gamma_3\}$. Vengono costruiti ad esempio 4 modelli che utilizzano le seguenti partizioni di Γ

	Γ'_i	Γ''_i
M_1	$\{\gamma_0\}$	$\{\gamma_1, \gamma_2, \gamma_3\}$
M_2	$\{\gamma_0, \gamma_2, \gamma_3\}$	$\{\gamma_1\}$
M_3	$\{\gamma_2\}$	$\{\gamma_0, \gamma_1, \gamma_3\}$
M_4	$\{\gamma_0, \gamma_1, \gamma_2\}$	$\{\gamma_3\}$

Si supponga che i risultati delle classificazioni di un nuovo record t siano $M_1(t) = \Gamma'_1$, $M_2(t) = \Gamma'_2$, $M_3(t) = \Gamma''_3$ e $M_4(t) = \Gamma'_4$. L'algoritmo procede sommando i punti come indicato qui sotto

γ_0	γ_1	γ_2	γ_3	
1	0	0	0	+
1	0	1	1	+
1	1	0	1	+
1	1	1	0	+
4	2	2	2	=

Si associa allora il record t alla classe γ_0. ▪

Manipolando l'algoritmo di learning : molti algoritmi di learning permettono di settare diversi parametri la cui modifica produce modelli diversi. Per implementare questa tecnica bisogna avere padronanza del particolare classificatore che si vuole modificare. Gli stessi alberi di decisione possono manipolati introducendo delle scelte random nella fase della loro induzione.

4.2 Bagging

Il metodo ensemble *Bagging* viene anche chiamato *Bootstrap Aggregating* e la struttura di questa tecnica è mostrata nell'algoritmo 4.2. L'ipotesi è di avere a disposizione un training set T con $|T| = n$ da poter manipolare.

Algoritmo 4.2 Bagging

```
1  Siano k gli insiemi di bootstrap da generare
2  for i ← 1 to k do
3        Crea un campione di bootstrap D_i di taglia n
4        Costruisci un modello M_i sul training set D_i
5  Classifica un nuova istanza t come M*(t) = Combine(M_1(t), ..., M_k(t))
   // δ(·) restiuisce 1 se l'argomento è vero e 0 altrimenti
```

Con il bagging si provvede a campionare ripetutamente il training set di input, con distribuzione di probabilità uniforme e con reinserimento, per costruire ad ogni iterazione un modello sul campione appena prodotto. Un campione di *bootstrap* è un campione casuale i cui elementi vengono estratti con reinserimento, per questo la tecnica prende anche nome di bootstrap aggregating. Tutti gli insiemi costruiti ad ogni iterazione sono di taglia uguale a quella del training set di partenza. Visto che si campiona con reinserimento, è quindi possibile che in uno stresso insieme siano presenti elementi ripetuti e che elementi del training set T vengano omessi. È utile far notare che in media, e quando gli insiemi sono abbastanza corposi, i training set costruiti contengono approssimativamente il 63.2% dei record di quello iniziale. Questo fatto è dimostrato nella proposizione 4.1. Una volta provveduto a costruire i k modelli un nuovo record viene classificato con la funzione Combine di cui si è discusso anche nell'introduzione ai metodi ensemble nella Paragrafo 4. Si ricorda che un modo semplice, e spesso utilizzato, per combinare le predizioni di più modelli è il sistema di voto.

Proposizione 4.1. *Dato un insieme di n elementi un campione di bootstrap B di taglia n, con n abbastanza grande, contiene in media il 63.2% degli elementi dell'insieme iniziale.*

Dimostrazione. Siano R e R_j con $1 \leq j \leq n$ due eventi così definiti:

$$R = \ r \text{ appartiene al data set di bootstrap } B$$

$$R_j = \ r \text{ viene aggiunto a } B \text{ alla } j\text{-esima estrazione}$$

è allora possibile calcolare la probabilità di R nel modo seguente

$$\begin{aligned}\mathbf{P}(R) &= \mathbf{P}\left(\bigcup_{j=1}^{n} R_j\right) = 1 - \mathbf{P}\left(\bigcap_{j=1}^{n} R_j^c\right) \\ &= 1 - \prod_{j=1}^{n} \mathbf{P}(R_j^c) = 1 - \prod_{j=1}^{n} \Big(1 - \mathbf{P}(R_j)\Big) \\ &= 1 - \left(1 - \frac{1}{n}\right)^n\end{aligned}$$

In queste derivazioni si utilizza il fatto che $R_{j'}$ e $R_{j''}$ sono indipendenti $\forall j', j''$ tali che $1 \leq j', j'' \leq n$. $\mathbf{P}(R_j)$ tende al valore approssimato di 0.632 per n grande

$$\mathbf{P}(R^\infty) = \lim_{n \to +\infty} 1 - \left(1 - \frac{1}{n}\right)^n = 1 - \frac{1}{e} \approx 0.632$$

È comodo costruire delle variabili aleatorie di Bernuolli in questo modo:

$$X_i = \begin{cases} 1 & \text{se } r_i = B \\ 0 & \text{se } r_i \neq B \end{cases}$$

Naturalmente, mantenendo l'ipotesi di n molto grande, $X_i \sim b(\mathbf{P}(R^\infty))\ \forall i:\ 1 \leq i \leq n$. Essendo tutte le variabili X_i i.i.d. è possibile dire che:

$$X = \sum_{i=1}^{n} X_i \sim Bin(n, \mathbf{P}(R^\infty))$$

X rappresenta *il numero di elementi non duplicati* nell'insieme di bootstrap B. Si dimostra la tesi calcolando il valore atteso di questa variabile aleatoria:

$$E(X) = n \cdot \mathbf{P}(R^\infty) = n \cdot 0.632$$

□

Il bagging migliora il generalization error rate riducendo la *varianza* del classificatore che indica qualitativamente quanto un classificatore è influenzato dalle perturbazioni del training set. Se il classificatore utilizzato è instabile, cioè è molto sensibile alle variazioni del training set, questa tecnica aiuta a ridurre gli errori dovuti alle fluttuazioni casuali dei dati di training. In ultima analisi, siccome ogni record ha la stessa probabilità di essere selezionato, il bagging è difficile che si concentri su una particolare istanza del training set. Grazie a questo fatto è molto meno sensibile all'overfitting.

In quanto al parametro k, che definisce il numero di modelli da costruire, la scelta è empirica. Nel lavoro originale di Breiman [21] i migliori risultati si sono ottenuti con $k = 10$.

4.3 Combine Multiple Models

In molti problemi di classificazione non è solamente importante che i modelli prodotti siano accurati ma vi è la necessità di poter osservare il meccanismo interno del processo per poter permettere agli utenti di utilizzare uno strumento comprensibile. Aggiungere trasparenza al modello, quindi non producendo modelli a scatola nera ma a *scatola bianca*, gli permette di guadagnare credibilità, facilita il processo di raffinamento della tecnica stessa e permette inoltre di avere il supporto degli esperti del campo in cui si sta applicando la classificazione.

È possibile ovviare, almeno in parte, al problema del rumore indotto dal training set con la tecnica del bagging, discussa nel paragrafo precedente, a discapito della trasparenza del modello. In [24] è proposta una tecnica generale per *combinare* più modelli di un metodo ensemble senza pregiudicare la chiarezza espressiva di un classificatore singolo. In linea di principio è possibile utilizzare un qualsiasi classificatore base e un qualsiasi metodo ensemble per produrre un modello che combina i vari prodotti.

L'idea è proposta nell'Algoritmo 4.3. Selezionato il numero di modelli k che si vogliono

Algoritmo 4.3 Combine Multiple Models

CombineMultipleModels($\mathcal{L}, \mathcal{E}, T, k, N$)

1 Costruisci k modelli utilizzando il metodo ensemble $\mathcal{E}$
  sul training set T e l'algoritmo di learning $\mathcal{L}$
2 $T' \leftarrow \emptyset$
3 **for** $i \leftarrow 1$ **to** N **do**
4 Prendi $t \in T$ random
5 $t[C] \leftarrow$ Combine($M_1(t), \ldots, M_k(t)$)
6 $T \leftarrow T \cup \{t\}$
7 Costruisci un modello M a partire dal training set T
  con l'algoritmo di learning $\mathcal{L}$
8 **return** M

far produrre al metodo ensemble $\mathcal{E}$ questi vengono utilizzati per produrre un training set artificiale di taglia $N + n$ dove n è la taglia del training set T fornito in input all'algoritmo e N è un altro parametro di input. Per costruire il nuovo insieme di learning vengono presi casualmente record da T ed aggiunti a T stesso modificandone la classe reale con quella predetta dalla classificazione combinata dei vari modelli ottenuti. Una volta composto il training set di taglia voluta lo si fornisce in input ad un classificatore per ottenere un nuovo modello che tiene conto dell'apporto di tutti quelli prodotti con il metodo ensemble selezionato. Il modello creato viene generato con l'algoritmo di learning $\mathcal{L}$ utilizzato anche nella produzione dei k modelli al passo 1 dell'algoritmo.

Nell'articolo in cui viene proposta questa tecnica viene testato il comportamento dell'algoritmo imponendo di utilizzare come learning algorithm *C4.5RULES*, che produce un

classificatore basato su regole a partire da un albero decisionale (per maggiori informazioni si legga [25]), e come metodo ensemble il bagging. In quasi tutti i casi si sono ottenute delle accuratezze più alte di quelle che si sono ottenute con un singolo classificatore e nella maggior parte dei casi i loro valori erano paragonabili a quelli ottenuti con il bagging. I migliori risultati si sono ottenuti settando k a 25, che è un valore più elevato di quello congeniale per il bagging, e N a 1000.

Capitolo 5

Alberi decisionali

In questo capitolo viene introdotta una tipologia importante di classification model: gli alberi decisionali. La definizione formale viene presentata nel Paragrafo 5.1 e il loro algoritmo di induzione è spiegato nel Paragrafo 5.3. Tale algoritmo fa pesantemente uso delle tecniche di split introdotte nel Paragrafo 5.2. Nel Paragrafo 5.4 viene spiegato rispettivamente in cosa consiste il fenomeno dell'overfitting: un fattore da tenere in considerazione per ottenere alberi accurati. Nel Paragrafo 5.5 viene evidenziato un difetto dell'algoritmo di induzione dovuto alla sua natura greedy. Per ultimo viene spiegato, nel Paragrafo 5.6, come si comporta C4.5, che è un'implementazione in C dell'algoritmo di induzione di alberi decisionali, con training set a dati mancanti. Tale trattazione è utile perché i modelli della risposta sostenuta alla terapia sono stati costruiti basandosi su un'implementazione alternativa di tale software.

5.1 Definizione

L'esistenza di questa tecnica è la dimostrazione di come si possa risolvere il problema della classificazione ponendo una serie di domande mirate sui valori degli attributi di un record di test. Ogni volta che si riceve una risposta viene posta la domanda successiva in modo che sia attinente al risultato ottenuto, il processo viene iterato finché non si ottiene la classe di tale record. La serie di domande, e le relative risposte, possono essere organizzate in una struttura ad albero. Famose implementazioni di algoritmi di induzione di alberi decisionali sono CART (*Classification And Regression Tree*) [26] e C4.5 [25].

Definizione 5.1. *Un albero decisionale o decision tree per un training set T è un albero in cui ogni nodo interno è associato a un test su uno degli attributi e gli archi verso i figli sono etichettati con i risultati distinti del test. Ogni foglia u è associata ad un sottoinsieme di record $T_u \subseteq T$ ed è etichettata con la classe di maggioranza dei record di T_u, inoltre i valori degli attributi dei record associati alla foglia u sono coerenti con i risultati dei test incontrati nel cammino dalla radice alla foglia u. Gli insiemi di record coperti dalle foglie creano una partizione del training set T.*

È possibile osservare che questo tipo di classificatori presentano i seguenti aspetti positivi:

- Sono di facile interpretazione;
- Ottengono una buona accuratezza su gran parte dei problemi reali di classificazione;
- Sono robusti rispetto al rumore e alla ridondanza tra gli attributi, cioè alla dipendenza totale o parziale tra alcuni degli attributi;
- Possono essere costruiti efficientemente.

In Figura 5.1 è rappresentato un decision tree costruito in base al training set fornito nell'Esempio 5.1.

Esempio 5.1. Si supponga che una compagnia di assicurazioni voglia identificare il legame che lega le classi di rischio, in cui vengono suddivisi i clienti, con l'età anagrafica e il tipo di vettura posseduta. Lo studio si deve basare su un gruppo di clienti già classificati nella corrispettiva classe. Il training set a disposizione è T, rappresentato qui sotto, e l'insieme delle classi è $\Gamma = \{A, B\}$.

Rid	Età	Tipo di Auto	Rischio
1	23	Berlina	A
2	18	Sportiva	A
3	43	Sportiva	A
4	68	Berlina	B
5	32	Furgone	B
6	20	Berlina	A

Tabella 5.1: Training set d'esempio per classificare clienti di una compagnia di assicurazioni in classi di rischio opportune

Un possibile albero di decisione è disegnato in Figura 5.1. Si noti come i nodi interni corrispondano a test sugli attributi e come i nodi foglia vengano etichettati con la classe di maggioranza per la rispettiva foglia. Si noti inoltre che i record di T vengono partizionati tra le foglie dell'albero. ▪

I test sugli attributi, effettuati nei nodi interni, differiscono a seconda del tipo di dati. La suddivisione del training set, dovuta ad esiti diversi nei test, è definita *split*. È possibile elencare due tipi di split a seconda che il tipo di attributo sia *nominale* o meno. L'esempio viene proposto per split binari. In caso di attributo nominale uno split binario sull'attributo A suddivide i record t con $t[A] \in \Phi$ e con $t[A] \notin \Phi$ dove Φ è un sottoinsieme del dominio di A. In caso che l'attributo non sia nominale, quindi categorico ordinale o di tipo numerico, lo split suddivide i record t con $t[A] \leq x$ da quelli con $t[A] > x$ dove x appartiene al dominio dell'attributo A. I corrispettivi split disegnati sull'albero di decisione sono rappresentati in Figura 5.2.

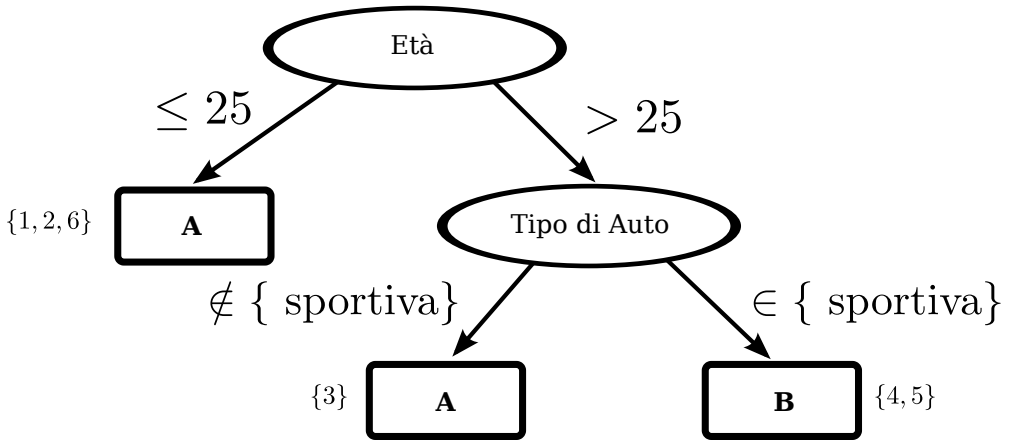
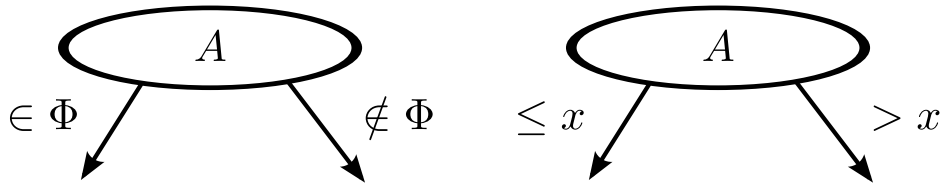


Figura 5.1: Decision tree d'esempio per una compagnia di assicurazioni



(a) Attributo nominale (b) Attributo non nominale

Figura 5.2: Split di tipo diverso

5.2 Selezione del miglior split

L'algoritmo di induzione di alberi decisionali costruisce un albero selezionando ad ogni nodo l'attributo che permette di ottenere lo *split* migliore ed inoltre specifica quale test deve essere effettuato su di esso. Dato un qualsiasi insieme di record E uno split è semplicemente una partizione indotta da un test su un attributo, ovvero:

$$E = E_1 \cup E_2 \cup \ldots E_k$$

La bontà dello split è legata alla differenza di *purezza* tra E e $(E_1, E_2, \ldots, E_k)$, dove la purezza indica quantitativamente se una classe è predominante sulle altre. Lo split deve far in modo che gli insiemi $(E_1, E_2, \ldots, E_k)$ siano globalmente più puri di E.

Ci sono varia misure di impurità che sono state utilizzate in letteratura, tutte hanno in comune che assumo valori elevati per insiemi di record la cui distribuzione di probabilità per le classi è uniforme e valori più bassi per distribuzioni non uniformi che indicano la predominanza di una determinata classe. Una generica misura di impurità $I(\cdot)$ viene quindi misurata su un insieme di record E.

Ricordando che $\Gamma = \{\gamma_0, \ldots, \gamma_{c-1}\}$ si definisce p_i come la frazione di record di E di classe γ_i:

$$p_i = \frac{|\{t \in E : t[C] = \gamma_i\}|}{|E|} \quad \text{con } 0 \leq i \leq c-1$$

Naturalmente

$$p_i \in [0,1] \ \forall i \ \text{ e } \sum_{i=0}^{c-1} p_i = 1$$

Misure di impurità più comuni sono:

(i) *Gini index*:

$$Gini(E) = 1 - \sum_{i=0}^{c-1} p_i^2$$

(ii) *Entropy*:

$$H(E) = -\sum_{i=0}^{c-1} p_i \log_2 p_i$$

(iii) *Classification error*:

$$Ce(E) = 1 - \max\{p_i : 0 \leq i \leq c-q\}$$

Assumono il valore 0 per distribuzioni empiriche di probabilità delle classi di E del tipo

$$p_i = \begin{cases} 1 & i = i^* \\ 0 & i \neq i^* \end{cases}$$

Per distribuzioni empiriche uniformi, cioè dove $p_i = \frac{1}{c}\ \forall i$, assumono il loro valore massimo che è rispettivamente $Gini(E) = 1 - \frac{1}{c}$, $H(E) = \log_2 c$ e $Ce(E) = 1 - \frac{1}{c}$.

Per capire quanto è efficace uno split è necessario mettere a confronto il grado di impurità del nodo padre e il grado di impurità dei nodi figli dopo lo split. Maggiore è la differenza migliore sarà la condizione di test usata per lo split. Supponendo di usare la generica misura di impurità $I(\cdot)$ si definisce tale differenza come *gain*:

Definizione 5.2. *Il gain per uno split del tipo $E \to (E_1, \ldots, E_k)$ è*

$$\Delta = I(E) - \sum_{j=1}^{k} \frac{|E_j|}{|E|} I(E_j)$$

Se si utilizza come misura di impurità l'entropia H il gain viene comunemente chiamato *information gain* e indicato con Δ_{info}. Purtroppo le misure di impurità appena introdotte tendono a favorire split che partizionano l'insieme di record dato in molte partizioni. Questo porta il modello ad adattarsi troppo bene al training set per poi non essere più in grado di rappresentare un altro data set sullo stesso insieme di attributi. I dati del

training set sono spesso affetti da rumore e non è sempre possibile fidarsi ciecamente di essi facendo in modo che il modello vi si adatti completamente, ma di questo si discute nella Paragrafo 5.4. Le metodologie di risoluzione sono principalmente due: impedire i test sui nodi interni a risposte multiple imponendo la costruzione di soli split binari o il pesare negativamente la costruzione di split a risposte multiple. Nel primo caso si va a ridurre il carico computazionale dell'algoritmo visto che si devono controllore solo tutti i possibili split binari. Una particolare strategia del secondo tipo è quella che sostituisce il gain al *gain ratio*. Questa è un particolare metodo utilizzato in C4.5 e visto che i modelli ottenuti in questa tesi vengono costruiti con una variante della sua implementazione vale la pena di essere citato. La definizione di gain ration è la seguente:

Definizione 5.3. *Il gain ratio per uno split del tipo* $E \to (E_1, \ldots, E_k)$ *è calcolato come*

$$GainRatio = \frac{\Delta_{info}}{SplitInfo}$$

dove

$$SplitInfo = -\sum_{j=1}^{k} \frac{|E_j|}{|E|} I(E_j)$$

Assume valore massimo quando tutte le partizioni hanno la stessa taglia, cioè $|E_j| = |E|/k \; \forall j$, ed è $\log_2 k$. Maggiore è il numero di insiemi in cui viene partizionato E (maggiore è k) maggiore sarà il termine al denominatore del gain ratio e quindi minore sarà il gain ratio stesso. Questo disincentiva la produzione di test con risultati multipli.

5.3 Algoritmo di induzione

In generale è possibile costruire un numero esponenziale di alberi decisionali dato un determinato insieme di attributi. La ricerca dell'albero con maggiore accuratezza è computazionalmente dispendiosa a causa della dimensione esponenziale dello spazio delle soluzioni. Il primo risultato, tra gli studi sulla complessità di questo problema, risale al 1976 in [27] e prova che la sola costruzione di un albero decisionale binario ottimo, dove l'albero ottimo è quello che minimizza il numero di test da effettuare per classificare un nuovo record, è *NP-Hard*. In [28] è possibile leggere un resoconto di quali sono i problemi legati alla costruzione di alberi decisionali che si dimostra appartenere alla classe *NP-Complete*. Ciò non preclude la possibilità di costruire algoritmi sub-ottimali per la risoluzione del problema *decision tree induction*.

In [29] è stata posta la base degli algoritmi attualmente utilizzati per la costruzione di alberi decisionali, ci si sta riferendo all'*algoritmo di Hunt*. Con il paradigma *greedy* si riescono ad ottenere alberi con buona accuratezza facendo una serie di decisioni locali ottime nella scelta dell'attributo che si utilizza per partizionare i dati. Naturalmente non è detto che queste scelte siano ottime anche globalmente. L'algoritmo di Hunt segue un approccio di tipo *Divide & Conquer*, mirando ad ottenere sottoinsiemi di record puri, cioè

afferenti alla stessa classe, ricorsivamente partizionando i training record. Lo pseudocodice per l'induzione di un albero decisionale è scritto nell'algoritmo in 5.1. Tale algoritmo necessita di due parametri di input: E come training set sullo schema di attributi $\mathcal{A} \cup C$ e appunto l'insieme di attributi $\mathcal{A}$ e la classe C.

Algoritmo 5.1 Algoritmo per decision tree induction

```
TreeGrowth(E, A, C)
 1  if StoppingCondition(E, A, C) then
 2      Crea una foglia leaf associata ad E
 3      leaf.label ← Classify(E, A, C)
 4  else
 5      Crea un nodo root
 6      root.test_cond ← FindBestSplit(E, A, C)
 7      V ← {v : v possibile risultato del test root.test_cond}
 8      for each v ∈ V do
 9          E_v ← {e ∈ E : root.test_cond(e) = v}
10          child ← TreeGrowth(E_v, A, C)
11          Aggiungi child come foglia di root con un arco etichettato v
12  return root
```

Le funzioni utilizzate sono:

- FindBestSplit$(E, \mathcal{A}, C)$: dato il sottoinsieme del training set E tenta di trovare il miglior split tra tutti i possibili split effettuabili sull'insieme di attributi $\mathcal{A}$. La qualità dello split può essere misurata, come indicato nelle sezioni precedenti, valutando la differenza tra la purezza dell'insieme E prima dello split e la partizione indotta dallo split.

- Classify$(E, \mathcal{A}, C)$: assegna una classe specifica all'insieme di record E. Solitamente questa funzione restituisce la classe di maggioranza di tale insieme valutata come

$$\arg\max_i \{p_i : 0 \leq i \leq c-1\} \text{ dove } p_i = \frac{|\{t \in E : t[C] = \gamma_i\}|}{|E|}$$

- StoppingCondition$(E, \mathcal{A}, C)$: restituisce il valore TRUE se è possibile terminare il processo ricorsivo e quindi non è più necessario partizionare l'insieme E, FALSE altrimenti. Un criterio utilizzato è quello di valutare se la taglia di E è al di sotto di una soglia fissata.

5.4 Overfitting

Oltre al banale fenomeno dell'*underfitting*, in cui il modello M non riesce a rappresentare bene nemmeno il training set e quindi a maggior ragione non rappresenterà bene l'insieme universo dei record U, bisogna gestire i casi di *overfitting*. Si può parlare di overfitting quando il modello rappresenta troppo accuratamente il training set E, ottenendo quindi un basso training error rate, ma ad esempio a causa della bassa cardinalità di E o al rumore dei record in esso contenuti l'accuracy è poco elevata. Si noti come è possibile espandere un albero decisionale affinché si adatti perfettamente ai dati di training per ottenere un training error rate di 0, in questo caso però il generalization error rate può essere grande mettendo in conto che potrebbero essere stati costruiti nodi che si adattano a record particolarmente rumorosi. Un nodo di questo tipo potrebbe abbassare la qualità del modello. Di contro un albero con un minore numero di nodi può avere un alto training error rate ma un basso generalization error rate.

Le principali cause dell'overfitting sono 3:

- La presenza di *rumore* nella classe target è determinante per la scelta di far adattare bene il modello al training set o meno. Bisogna distinguere i casi in cui la classe è *intrinsecamente rumorosa* e in cui il rumore è *indotto dal training set* utilizzato per costruire il modello. Nel primo caso pur avendo a disposizione grossissime quantità di dati sul problema in esame è possibile che il classificatore non riesca ad abbassare il generalization error rate al di sotto di un limite minimo. In ultima analisi la classe può anche non essere definita deterministicamente, è possibile cioè che esistano record con lo stesso valore di attributi ma etichettati con classi diverse. Nel secondo caso è determinante la scelta degli attributi che si vogliono legare alla classe: ad esempio, ipotizzando di voler studiare il grado di istruzione di una popolazione con delle tecniche di classificazione utilizzare l'altezza di un individuo come parametro discriminante è chiaro che è una scelta discutibile. Scegliere degli attributi che non sono legati alla classe stabilisce a priori che il training set estratto sarà rumoroso. Sempre relativamente alla seconda categoria di rumore, quello da imputare al training set, fa parte il rumore introdotto da record i cui valori degli attributi sono affetti da incertezza. In questo caso sebbene l'attributo stesso sia legato significativamente alla classe target un modello che si adatti completamente a tale training set non permette di ottenere buone prestazioni.

- La bassa cardinalità del training set può indurre ad overfitting: i pochi dati a disposizione potrebbero non permettere di formulare ipotesi significative sul legame tra gli attributi e la classe.

- Fenomeni legati a *multiple comparison procedure*. Maggiore è il numero di possibili split che possono essere effettuati più è facile che venga aggiunto uno split spurio, cioè uno split che diminuisce l'impurezza media dei nodi figli rispetto a quella del nodo padre solo per i record del training per cui è stato scelto ma non per record qualsiasi. Per un algoritmo di tipo greedy con un numero di attributi elevato diventa

anche più grande lo spazio delle soluzioni ed è quindi più probabile imbattersi in un ottimo locale invece di uno globale.

In particolare quando il classification model è un albero decisionale è possibile ovviare in parte a questo problema con il *pruning*, o potatura, dell'albero stesso. Viene definita *prepruning* la tecnica che vuole evitare l'overfitting fermando il processo di costruzione del decision tree prima che si adatti in modo ottimale al training set, si parla di *postpruning* quando una volta costruito l'albero decisionale completo vengono contratte parti di questo con approccio bottom-up.

Ad esempio la tecnica di pruning utilizzata da C4.5 prende nome di *error-based pruning*, la decisione di potare o meno parti dell'albero vengono prese in base ad una stima pessimistica del generalization error a partire dal training error.

Dato un nodo foglia che copre N record del training set, di cui K classificati erroneamente, il valore $\frac{K}{N}$ può essere visto come una stima del contributo all'errore generalizzato apportato da quella foglia. Prendendo l'estremo superiore dell'intervallo di confidenza di questa stima C4.5 inferisce su quello che potrà essere il generalization error in un caso pessimistico. Quello che si sta facendo è sovrastimare l'errore generalizzato apportato da una foglia a partire dal valore $\frac{K}{N}$. Naturalmente l'ampiezza dell'intervallo viene gestita impostando un livello di confidenza α. Un sottoalbero viene potato se peggiora la stima del generalization error rate rispetto a quella che si otterrebbe se si sostituisse il sottoalbero con una sola foglia.

Un altra tecnica di pruning, proposta in [30], sfrutta un approccio dell'*Information theory*. Il modello M che viene costruito deve minimizzare la somma di costi: $Costo(E|M)$ che è funzione del numero di training error e $Costo(M)$, cioè il costo del modello stesso. Ad esempio è possibile associare un costo che aumenta proporzionalmente al numero di nodi dell'albero decisionale. All'aumentare della complessità del modello aumenta $Costo(M)$ ma diminuisce $Costo(E|M)$ visto che probabilmente il numero di training error diminuirà. Viceversa diminuendo la complessità del modello diminuisce il primo fattore ma aumenta il secondo. Trovare il giusto trade-off è il punto chiave di questa tecnica definita *Minimum Description Length Principle* (MDLP).

5.5 Split su più attributi

In questo paragrafo si vuole parlare brevemente dei principali punti deboli nel meccanismo di classificazione degli alberi decisionali fornendo un esempio. I tipi di test fatti ai nodi interni dell'albero permettono solo di suddividere lo spazio del dominio in iper-rettangoli m-dimensionali, dove $|\mathcal{A}| = m$. Non tutti i problemi permettono di essere risolti con elevata accuratezza con questo approccio. Un altro punto debole è naturalmente la strategia greedy. Basti pensare che solo imponendo split su *coppie* di attributi il risultato potrebbe in certi casi essere notevolmente migliore.

Si pensi di voler trovare lo split ottimo su un unico attributo rispetto al dataset E rappresentato in Figura 5.3, esempi di questo tipo sono chiamati *XOR-like problems*.

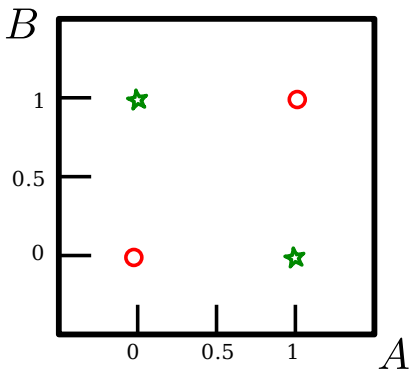


Figura 5.3: Dataset E di un problema XOR-like

L'insieme degli attributi in esame è $\mathcal{A} = \{A, B\}$ e si vuole usare come misura di impurità $I(\cdot)$ il *Gini index*.

Una volta calcolato il Gini index del dataset E

$$\text{Gini}(E) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

è possibile tentare di trovare un valore di cutoff o per l'attributo A o per l'attributo B, ma comunque lo si scelga il *gain* che si ottiene è nullo. Di seguito si sviluppano per rendere chiare le idee i calcoli relative ad uno split su A, come in Figura 5.4.

$$\text{Gini}(E_1) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2} = \text{Gini}(E_2)$$

Quindi il gain risultante è

$$\Delta = \frac{1}{2} - \left[\left(\frac{1}{2}\right)\frac{1}{2} + \left(\frac{1}{2}\right)\frac{1}{2}\right] = 0$$

Imponendo uno split su entrambi gli attributi ad esempio $A \leq 0.5 \ \wedge \ B \leq 0.5$ come in Figura 5.5 cambierebbe il Gini index di E_1 e E_2 in questa maniera

$$\text{Gini}(E_1) = 1 - 1 = 0$$

$$\text{Gini}(E_2) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

in modo da produrre un gain di

$$\Delta = \frac{1}{2} - \left[(0)\frac{1}{4} + \left(\frac{4}{9}\right)\frac{3}{4}\right] = \frac{1}{3}$$

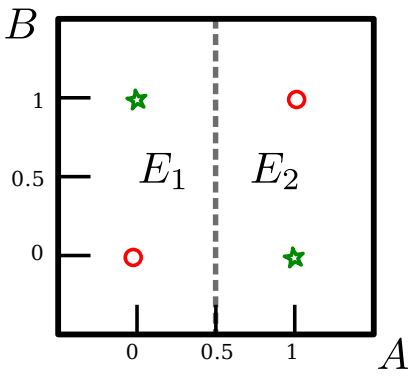


Figura 5.4: Split sull'attributo A

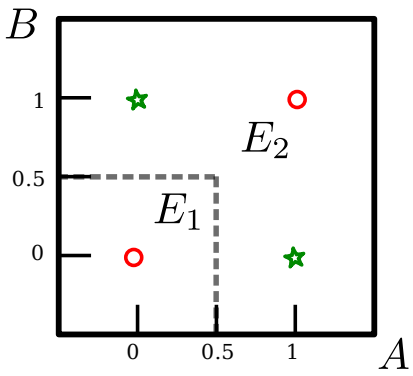


Figura 5.5: Split sia su A che su B

Certo è possibile separare il dataset E, sempre linearmente, in un modo migliore (Figura 5.6), ma si dovrebbe introdurre una disequazione lineare per renderlo possibile peggiorando la chiarezza nella lettura dello split stesso.

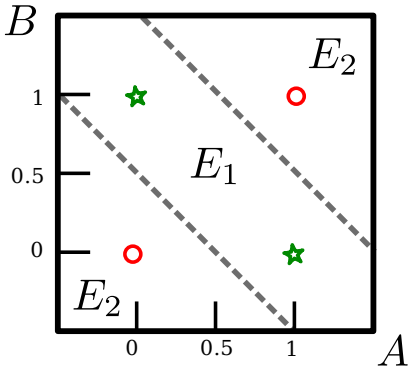


Figura 5.6: Split migliore del dataset E

Se l'algoritmo di induzione di alberi decisionali fosse messo alla prova su questo esempio non produrrebbe nessun split tentando di osservare attributo per attributo singolarmente. Se invece la ricerca venisse fatta a coppie di attributi certamente lo split relativo alla Figura 5.5 verrebbe preso in considerazione. Questo difetto è dovuto alla natura *greedy* dell'approccio di induzione dei decision tree.

5.6 Dati mancanti

Citando Quinlan nella trattazione di C4.5: "È un fatto sfortunato della vita che vi siano dati con valori mancanti negli attributi". Questo può accadere a causa di vari fattori quali ad esempio la difficoltà o il costo nel misurare quel particolare valore. Capita spesso in ambito medico che certi esami non vengano fatti perché non di routine. Una possibilità è quella di utilizzare solo parametri il cui valore è noto per tutte le istanze in studio, ma così facendo non si utilizzerebbe una buona parte dei dati a disposizione. Si vuole pertanto

presentare brevemente come viene risolto il problema dei valori mancanti degli attributi in C4.5 [25], visto che per questa tesi è stata utilizzata una sua implementazione.

Dato un insieme di record di training E l'approccio di C4.5 con i valori mancanti in un attributo A è il seguente. La produzione degli split relativi a tale attributo è fatta contando solo le istanze di E a valori noti. Una volta prodotto lo split $E \to (E_1, \ldots, E_k)$ il suo valore informativo viene ridotto della frazione f di record a valore nullo:

$$\Delta_{info} = f \cdot \left(H(E) - \sum_{j=1}^{k} \frac{|E_j|}{|E|} H(E_j) \right)$$

Inoltre lo *SplitInfo* viene calcolato come se vi fossero $k+1$ sottoinsiemi, dove viene contato anche quello a valori mancanti per A.

Una volta prodotto lo split le istanze a valore nullo in A vengono associate ad ogni E_i con peso diverso. Tale peso è rappresentativo della probabilità di tale record di appartenere ad un dato sottoinsieme. Supponendo che v_i con $1 \leq i \leq k$ siano le uscite del test sull'attributo A che genera lo split, il peso di un record t, con $t[A]$ non noto, per il sottoinsieme E_i si calcola:

$$w_i = \widehat{\mathbf{P}}(t[A] = v_i) = \frac{|\{t \in E : t[A] = v_i\}|}{|E|}$$

Naturalmente ad un certo punto della computazione è possibile che un'istanza t abbia già un peso non unitario assegnato, essendo E una partizione di un insieme più grande. In questo caso si procede considerando l'istanza con peso frazionario anche nel calcolo delle misure di impurità. Se un record t deve essere assegnato nuovamente a tutti i sottoinsiemi E_i di un'ulteriore split su un attributo B (quindi quando $t[B]$ è sconosciuto) si procede all'aggiornamento del peso così:

$$w_i \leftarrow w_i \cdot \widehat{\mathbf{P}}(t[B] = v_i) = w \cdot \frac{|\{t \in E : t[B] = v_i\}|}{|E|}$$

Questo metodo è un buon metodo per tenere conto anche degli attributi a valori mancanti, salvando così una buona percentuale di dati reperiti dall'operatore. Inoltre, la somma della cardinalità degli insiemi associati ai nodi foglia da la cardinalità del training set come accade per un albero indotto per un training set completo (senza valori mancanti).

Un approccio simile viene utilizzato per classificare un record sconosciuto. Se si incontra un nodo associato ad un test la cui risposta non può essere determinata vengono esplorate tutte le possibili uscite di tale test. Essendoci la possibilità quindi di avere cammini multipli dalla radice a varie foglie per una singola istanza i risultati vengono combinati aritmeticamente. Ottenuta la distribuzione di probabilità per le classi del record t, con l'apporto di tutte le foglie in cui potrebbe finire tale record, questo viene etichettato con la classe più probabile.

Capitolo 6

Probability Estimation Tree (PET)

In questo capitolo vengono introdotti, nel Paragrafo 6.1, gli alberi per la stima di probabilità, per fornire un approccio alternativo alla classificazione. La loro induzione è descritta nel Paragrafo 6.2. Le tecniche per valutare le loro prestazioni sono reperibili invece nel Paragrafo 6.3: tale paragrafo presenta le ROC curve, metodo utilizzato per valutare i modelli predittivi ottenuti in questa tesi.

6.1 Definizione

Etichettare record la cui classe non è nota con il valore predetto da un modello di classificazione spesso non è abbastanza. In molte applicazioni è più importante fornire la *probabilità* di appartenenza a una classe per un dato record. Questo fatto è particolarmente rilevante in medicina: è difficile discriminare esattamente se un paziente guarirà o meno. Etichettare una persona come possibile *responder* ad una terapia in base alla predizione di un modello di classificazione può portare ad ottenere alte frequenze di falsi posivi e alte frequenze di falsi negativi. È più utile conoscere la probabilità di essere un responder in modo che il medico possa sapere quanto è verosimile che tale paziente riesca a guarire. Applicazioni che richiedono l'approccio appena introdotto mirano a discriminare i record positivi da quelli negativi con le probabilità: tale obiettivo viene definito ordinamento dei record in base alle probabilità.

In letteratura ci si riferisce agli alberi per la stima di probabilità con l'acronimo PET (*Probability Estimation Tree*). Essi sono utili per per l'ordinamento basato sulle probabilità o *probability-based ranking*, in più mantengono le stesse caratteristiche degli alberi decisionali definiti nel Capitolo 5 come la comprensibilità, l'accuratezza e l'efficienza della loro costruzione su dataset di grandi dimensioni. Per formalizzare, viene proposta la loro definizione:

Definizione 6.1. *Un Probability Estimation Tree (PET) per un training set T è un albero decisionale dove in ogni nodo foglia u è presentata la distribuzione di probabilità relativa alla classe dei record T_u associati ad essa.*

Spesso, a differenza di quanto dice la Definizione 6.1, anche nei nodi interni di un PET viene rappresentata la distribuzione di probabilità dei record ad essa associati. È doveroso sottolineare che la stima di probabilità di appartenenza ad una classe per un nuovo record è data solo dalle foglie, quindi indicare la distribuzione ai nodi interni può essere considerata solo un'informazione per l'utente che osserva il PET, ma tale stima non viene validata nella validazione complessiva del modello. Un esempio di PET sulla stima di probabilità di guarigione è mostrato in Figura 6.1. Si ipotizzi che le possibili classi siano solo due: chi guarisce (paziente "Guarisce") e chi non guarisce (paziente "Non guarisce") . Per rappresentare le distribuzioni di probabilità relative ai nodi dell'albero in modo chiaro vengono prodotti dei diagrammi a torta. La probabilità segnata a sinistra dei nodi è quella di non non guarire, a destra invece quella di guarire. Sopra ogni nodo è presentata la cardinalità del sottoinsieme di record a lui associato. Si noti come i nodi interni siano associati ad un test su un attributo. Nell'esempio proposto si può vedere come i maschi abbiano meno pro-

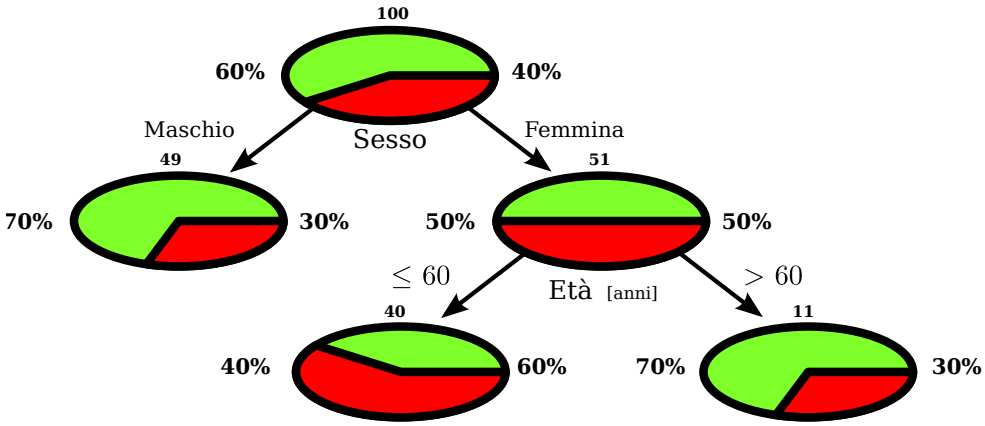


Figura 6.1: PET d'esempio sulla probabilità di guarigione

babilità di guarigione delle femmine e che le ultime possano essere suddivise ulteriormente in base all'età. I diagrammi a torta, dove in rosso è rappresentata la classe "Guarisce" e in verde chiaro la classe "Non guarisce", rendono immediata la comprensione del grafico.

6.2 Induzione di PET

I metodi per l'induzione di (PET) sono stati inizialmente studiati da Provost in [31]. La struttura base è quella descritta nell'Algoritmo 5.1 TreeGrowth con una modifica sostanziale: la chiamata a funzione Classify in riga 3 viene sostituita con la funzione EstimateClassProbabily. Essa prende in input E, $\mathcal{A}$ e C che sono rispettivamente l'insieme di training, l'insieme degli attributi e la classe C. Naturalmente E deve essere definito sullo schema $\mathcal{A} \cup C$. La sua implementazione più spontanea, analizzata sempre nel lavoro di Provost, è la stima delle probabilità delle classi con il metodo di *massima verosimiglianza* (per un'analisi specifica, le stime di massima verosimiglianza si possono trovare in [14]). La probabilità che un record t appartenga alla classe γ trovandosi in una

foglia a cui è associato un sottoinsieme E è così stimata:

$$\widehat{\mathbf{P}}_M(t[C] = \gamma) = \frac{|\{r \in E : \ r[C] = \gamma\}|}{|E|}$$

Visto che viene utilizzato l'approccio classico o Fisheriano tale parametro è un *numero* non noto e non una variabile aleatoria. Non si introduce pertanto nessun intervallo di confidenza di tale stima. Il difetto di questa tecnica è che, come è logico intuire, una stima fatta su insieme E piccolo risulta poco accurata.

Per migliorare la stima di probabilità di appartenenza alla classe in insieme E piccoli viene utilizzata la *correzione di Laplace*. Tale metodo permette di non utilizzare l'approccio Bayesiano (che considera il parametro da stimare una variabile aleatoria) e di migliorare la stima nel caso di campioni piccoli. La probabilità che un record t appartenga alla classe γ data una foglia a cui è associato E è:

$$\widehat{\mathbf{P}}_L(t[C] = \gamma) = \frac{|\{r \in E : \ r[C] = \gamma\}| + 1}{|E| + c}$$

Dove c è il numero delle classi studiate. Si veda l'Esempio 6.1 per capire come questa stima non permetta di ottenere valori estremi di probabilità per piccoli campioni, a differenza del metodo della massima verosimiglianza.

Esempio 6.1. Sia E un insieme di training sullo schema di attributi $\mathcal{A} \cup C$. Le possibili classi C sono 2 ($c = 2$) e l'insieme è composto da 5 elementi $|E| = 5$. Tutti gli elementi di E appartengono ad una sola classe, sia essa γ_0. Il metodo della massima verosimiglianza stimerebbe la probabilità di appartenere a γ_0 per un record che finisce in una foglia u associata ad E con

$$\widehat{\mathbf{P}}_M(t[C] = \gamma_0) = \frac{|\{r \in E : \ r[C] = \gamma_0\}|}{|E|} = \frac{5}{5} = 1$$

Invece, introducendo la correzione di Laplace, tale valore verrebbe:

$$\widehat{\mathbf{P}}_L(t[C] = \gamma_0) = \frac{|\{r \in E : \ r[C] = \gamma_0\}| + 1}{|E| + c} = \frac{5+1}{5+2} \approx 0.86$$

L'ultimo metodo permette di ottenere valori meno estremi di stima di probabilità. ▪

Intuitivamente un PET non sembra un metodo adatto alla stima di probabilità come lo potrebbe essere uno numerico che stimi le probabilità delle classi fornendo un output nell'intervallo $[0, 1]$. Gli alberi possono invece stimare le probabilità con un qualsiasi grado di precisione. È infatti ovvio che un albero possa stimare la distribuzione di probabilità di un attributo categorico, basta fornire uno split che suddivida il training set nei vari valori di tale attributo. Anche per gli attributi continui un PET di dimensione adeguatamente grande può stimare qualsiasi distribuzione di probabilità delle classi. Si osservi la Figura 6.2 dove è rappresentata la distribuzione di probabilità di una classe per l'attributo continuo

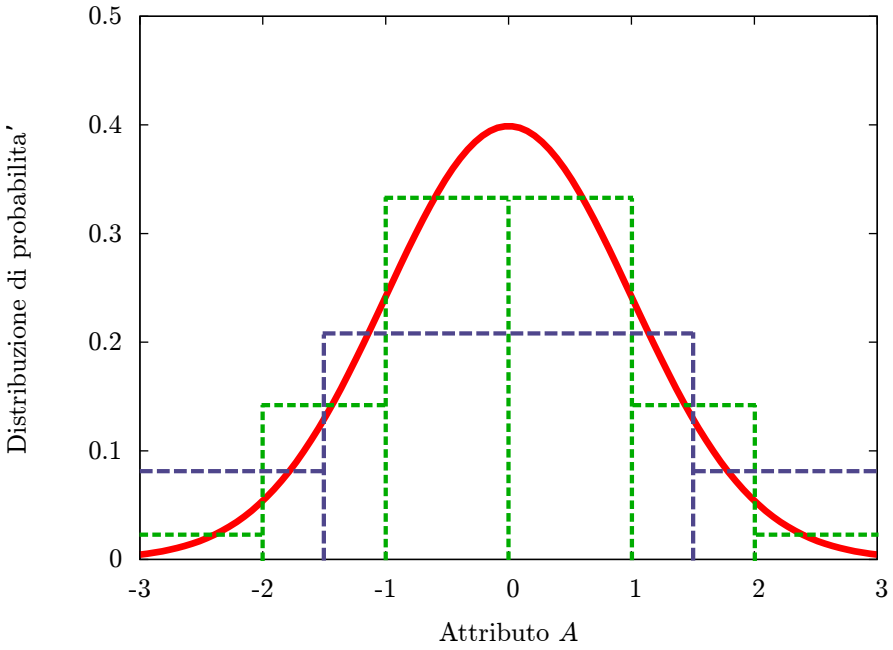


Figura 6.2: Distribuzione di probabilità per l'attributo continuo A

A. Fornendo uno split ternario che suddivide i record t con $t[A] \in (-\inf, -1.5]$, $t[A] \in (-1.5, 1.5]$ e $t[A] \in (-1.5, +\inf)$ si può stimare la distribuzione di probabilità di A ad un certo livello di dettaglio (linea blu), ma aumentando la cardinalità dello split o inducendo split consecutivi è possibile stimare tale probabilità a grana più fine. Si intuisce che il numero di istanze del training set deve essere molto elevato per poter spingersi ad un buon dettaglio nella stima (linea verde).

Una volta capito che *non* è dovuta alla struttura degli alberi la cattiva qualità delle stime per PET costruiti con gli algoritmi di induzione classici, nel lavoro di Provost [31] viene evidenziato che la falla è dovuta alle tecniche di pruning. La causa risulta chiara osservando le distribuzioni di probabilità in Figura 6.3 di un attributo A in due classi. L'albero di decisione più piccolo che massimizza l'accuratezza è un albero con un solo nodo interno che induce uno split binario per i record con $A \leq 0$ e $A > 0$. L'albero più piccolo è buono perché cerca di minizzare l'overfitting come spiegato nel Paragrafo 5.4. Ma così facendo, non permette di stimare accuratamente la distribuzione di probabilità dell'attributo. Visto che con le tecniche di pruning si tenta di ridurre al minimo l'altezza dell'albero sono proprio loro da imputare al cattivo comportamento dei decision tree classici per la stima di probabilità.

6.3 Metriche di prestazione

È ovvio che non sia possibile utilizzare per i PET le stesse metriche di prestazione del Paragrafo 3.2 specifiche per la classificazione. In questo paragrafo si vuole introdurre brevemente la *ROC curve* per poter valutare la prestazione di un PET.

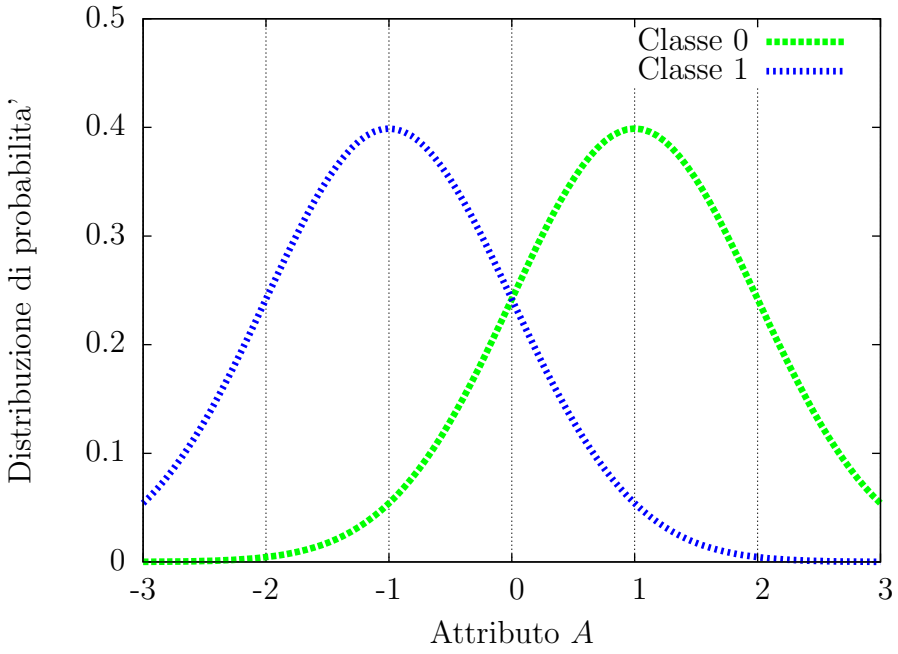


Figura 6.3: Distribuzione di probabilità dell'attributo A in due classi

La ROC curve, che è l'acronimo di *Receiver Operating Characteristic curve*, è ampiamente utilizzata in ambito medico e solo recentemente è stata introdotta nell'ambito del *Machine Learning*. Viene utilizzata per studiare la qualità di classification technique la cui classificazione è *binaria*. La sua applicazione più comune è quella di studiare modelli, quali ad esempio le reti neurali (cfr. [17], Paragrafo 5.4), il cui output è un valore di probabilità. Il suo utilizzo per valutare i PET per stimare la probabilità di appartenere ad una data classe diventa quindi naturale.

Viene considerato insieme di istanze la cui classe (binaria) è nota, sia esso E. Come è consueto fare per i problemi di classificazione binari, anche questo paragrafo una classe viene indicata come positiva e l'altra come negativa. Ad ogni istanza di E il PET assegna un valore di probabilità di appartenere alla classe dei positivi. La ROC curve è un approccio grafico per valutare la capacità del modello di ordinare le istanze positive rispetto a quelle negative con le probabilità. Un valore di probabilità deve essere scelto come soglia per considerare positive le istanze con probabilità sopra di essa e negative le altre. Ogni punto nella ROC curve è ottenuto calcolando il FPR (*False Positive Rate*) e il TPR (*True Positive Rate*) selezionando una data soglia. La loro formula viene qui richiamata:

$$FPR = \frac{FP}{TN + FP}$$

$$TPR = \frac{TP}{FN + TP}$$

Tali valori variano nell'intervallo [0,1]. Se il modello è un buon modello, a valori alti di probabilità corrispondono istanze positive e a valori bassi istanze negative. Variando la

soglia di probabilità scelta da 0 ad 1 e calcolando i valori TPR e FPR corrispondenti si vorrebbero ottenere punti che hanno un valore alto del primo al variare del secondo. Si ipotizzi di partire dalla soglia di probabilità 0: tutte le istanze vengono identificate come positive perciò nello spazio della ROC curve ($[0,1]^2$) si disegna un punto in $(1,1)$. Tutti i positivi saranno predetti come positivi, quindi $TPR = 1$, ma anche tutti i negativi saranno predetti come positivi, quindi $FPR = 1$. Alzando la soglia di probabilità, se l'ordinamento è coerente con le classi, il valore FPR diminuirà visto che i negativi hanno probabilità basse. Il TPR invece rimane ad 1 finché non vi è alcun vero positivo al di sotto della soglia scelta. L'ultimo punto disegnato con questo procedimento è il punto $(0,0)$ per una soglia di probabilità fissata a 1. Questo caso è l'opposto di quello iniziale: tutti i negativi sono predetti come negati e tutti i positivi sono predetti come negativi. Un esempio di ROC curve è mostrato in Figura 6.4. È possibile analizzare qualitativamente una ROC curve

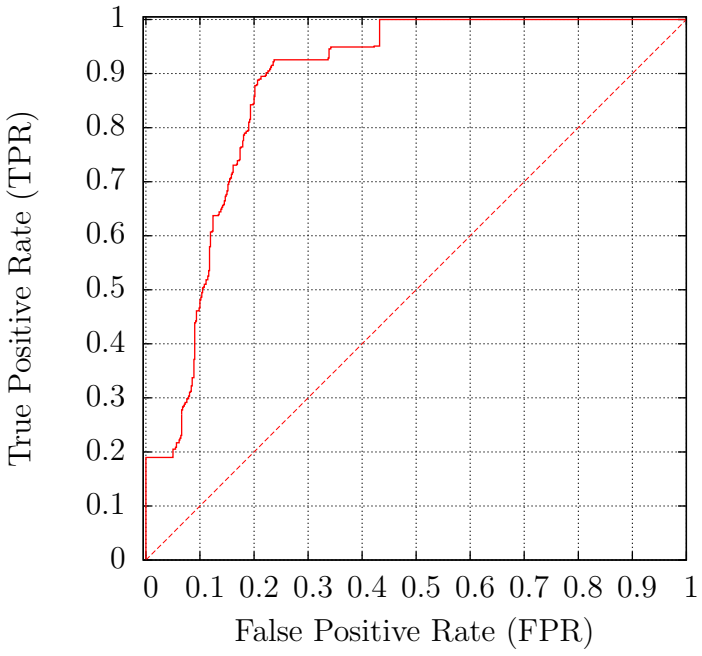


Figura 6.4: Esempio di ROC curve

osservando quanto sia schiacciata all'asse delle ordinate o alla retta definita da $TPR = 1$. Il primo caso indica che il modello riesce ad identificare le istanze negative associandovi basse probabilità, il secondo che riescono a venire identificate le istanze positive e ad esse si associano probabilità elevate. Non è detto che si riescano ad ottenere buoni risultati per entrambe le classi. La ROC curve è utile strumento per valutare la qualità di un modello in presenza di sbilanciamento tra le classi.

Per concludere si specifica che un valore riassuntivo della capacità del modello di ordinare i record in base alla probabilità delle classi è l'area sotto la ROC curve, definita AUC cioè *Area Under (ROC) Curve*. Il valore massimo che può assumere è naturalmente 1, il minimo è 0. Una curva è migliore mediamente se rimane il più possibile schiacciata

all'angolo sinistro e in alto del grafico. Ed esempio il valore di AUC del curva in Figura 6.4 è 0.858.

6.3.1 ROC curve con cross-validation

Per valutare la ROC curve di un PET è bene non utilizzare lo stesso insieme di training da cui è costruito. È possibile utilizzare un insieme di validazione distinto per costruirla ma una stima migliore della ROC ottenuta da un PET si ottiene con la k-fold cross-validation. Tale tecnica prevede che venga costruito un insieme E' la cui classe è nota e alle cui istanze è associato un valore di probabilità per ottenere la ROC curve. Il tutto viene fatto partendo da un training set E. Questo viene partizionato in k sottoinsiemi E_i di taglia $|E|/k$. L'insieme E' viene quindi ottenuto in k passi la cui i-esima iterazione prevede che venga costruito un PET su $E \setminus E_i$, associato il valore di probabilità alle istanze di E_i con tale modello e che queste vengano aggiunte ad E'. Con l'insieme E' viene disegnata la ROC curve relativa alla k-fold cross-validation. Le ROC curve ottenute con questo metodo non sono ottimistiche come quelle costruite sullo stesso training set utilizzato per fare il PET.

A seconda di come viene randomizzato il training set è possibile ottenere ROC curve diverse anche con la cross-validation appena introdotta. Pertanto, sarebbe bene avere una ROC curve mediata rispetto a quelle ottenute con diverse randomizzazioni dello stesso training set. In [32] sono proposti i metodi allo stato dell'arte per mediare più ROC curve. Quello più semplice stabilisce di utilizzare gli insiemi da cui si vogliono ottenere le ROC da mediare e fonderli in uno. Per ogni istanza dell'insieme ottenuto si ha la classe, positiva o negativa, ed un valore di probabilità associato. Tale insieme può essere utilizzato per costruire una ROC curve unica che racchiude le informazioni di tutte le m che si volevano mediare.

È pertanto possibile ottenere la ROC curve media di m curve ottenute con una k-fold cross-validation con questa metodica.

Capitolo 7

Analisi del problema clinico

In questo capitolo viene introdotto il problema medico, oggetto di questa tesi, nel Paragrafo 7.1. I motivi per cui è necessario predire la risposta sostenuta alla terapia vengono elencati analizzando la popolazione in studio nel Paragrafo 7.2. Dopodiché, nel Paragrafo 7.3, viene ripercorso l'iter logico che ha permesso di arrivare alla soluzione adottata, la cui presentazione è proposta in dettaglio in Paragrafo 7.4.

7.1 Introduzione

L'epatite C è una malattia lentamente ma inesorabilmente progressiva che determina una cronicizzazione nell'80% dei casi. In Italia, l'epatopatia HCV correlata provoca la morte di migliaia di individui ogni anno ed è la principale causa di tumore primitivo del fegato e trapianto epatico. La terapia antivirale può garantire un reale vantaggio clinico nei soggetti che eradicano l'infezione definitivamente, con risvolti anche economici per gli alti costi sostenuti per questo tipo di patologia.

La terapia per l'epatite C non è efficace in tutti i casi: purtroppo il 40% dei soggetti non è in grado di eradicare completamente il virus dall'organismo. Questa condizione di resistenza alla terapia antivirale è dovuta a molti fattori, tra i quali giocano un ruolo rilevante il genotipo e l'attività replicativa del virus. L'efficacia della terapia quindi si misura con la capacità di sopprimere la replicazione virale (cioè la viremia). Il soggetto che eradica l'infezione a fine terapia è detto "Responder"; se mantiene questa risposta anche a 6 mesi dalla sospensione della terapia è detto invece "Long Term Responder" (LTR). Purtroppo in alcuni casi si assiste alla temporanea soppressione della viremia (Responder), ma dopo la sospensione della terapia (entro 6 mesi) si presenta una riattivazione: questi soggetti si definiscono "Relapser" (RR). I rimanenti casi trattati mostrano un profilo di alta resistenza durante la terapia e restano sempre viremici, questi casi, detti "Non Responder" (NR), devono essere rapidamente identificati e sottratti al trattamento inefficace. I non responder e i relapser sono soggetti sfavoriti da molti punti di vista:

1. Non guariscono dalla malattia perché non eradicano l'infezione con la terapia;

2. Subiscono una terapia che peggiora notevolmente la qualità di vita e sono esposti a numerosi, anche gravi, effetti collaterali senza alcun beneficio;

3. Sono partecipi di inutili e non indifferenti spese per il sistema sanitario.

Si noti che la predizione della risposta alla terapia dell'epatite C è molto importante dal punto di vista farmaco-economico: si può notevolmente migliorare il rapporto di costo-efficacia del trattamento, evitando la terapia nel caso dei resistenti e garantendo l'adeguato completamento del ciclo di terapia nei casi con reale possibilità di guarigione.

Il virus dell'HCV è presente in natura in 4 genotipi principali. Ormai è accertato che i genotipi 1 e 4 siano quelli più resistenti alla terapia. Le linee guida internazionali propongono di far durare la terapia standard per 12 mesi nel genotipo 1 e 4 e per 6 mesi nel genotipo 2 e 3. Si osservi in Figura 7.1 la percentuale di pazienti che ottengono la risposta sostenuta in relazione rispettivamente al genotipo 1 o 4 e al genotipo 2 o 3; nei primi il 44.6% dei soggetti riescono ad essere curati, cioè eradicano completamente l'infezione (LTR), mentre nei secondi tale percentuale sale all'81.1%. Questa differenza è molto significativa.

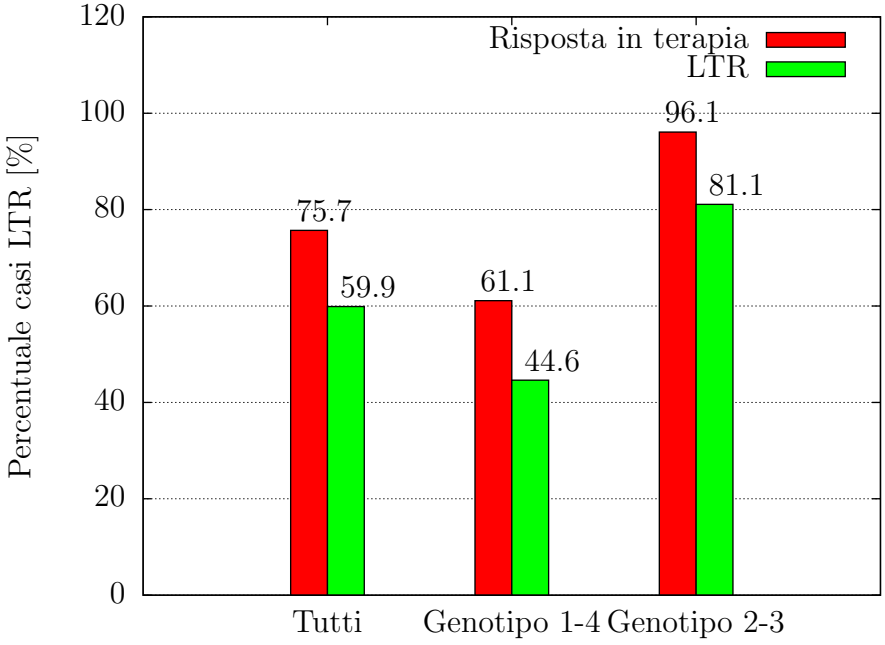


Figura 7.1: Percentuale di pazienti che ottengono la risposta in terapia e la riposta sostenuta in base al genotipo

In questo lavoro si sono analizzati i pazienti affetti da epatite C con genotipo 1 o 4, nei quali la predizione sulla risposta alla terapia è cruciale a causa dell'elevatissima percentuale di NR e RR. È apparso logico concentrare gli sforzi su questi soggetti perché più numerosi epidemiologicamente e perché in questa classe gli aspetti di costo-efficacia sono più sensibili.

In questa tesi ci si è affidati alle tecnologie sviluppate per la classificazione per poter ottenere l'analisi di tipo statistico di un campione significativo di pazienti con risposta alla terapia ben definita (casi NR, RR e LTR). In particolare, si è sviluppato un metodo computazionale che produce un modello semplice ed efficace per predire la risposta sostenuta di un individuo alla terapia. Tale modello viene costruito in base a pazienti il cui decorso terapeutico è già noto e può essere utilizzato per predire la risposta sostenuta di un nuovo paziente nelle varie fasi della terapia: basale o pretrattamento, al primo e al terzo mese durante il trattamento. Questi istanti temporali sono cruciali per predire l'esito della terapia come verrà spiegato nel prossimo paragrafo, in quanto delineano i vari profili di pazienti che sono suscettibili di una più rapida o più tardiva risposta alla terapia.

I modelli costruiti permetteranno allo specialista di soffermarsi ad osservare i parametri clinici di un nuovo paziente che meglio rappresentano l'andamento della terapia fino a quell'istante. Si potrà dunque assegnare automaticamente una probabilità di riuscita al trattamento, estrapolando inoltre un comportamento futuro. Il valore dei parametri selezionati dal modello sono facilmente misurabili con test di routine: questa è una caratteristica essenziale per uno strumento realmente utilizzabile in pratica. Visto la natura altamente dinamica del problema, ci si è spinti oltre la semplice predizione. I modelli costruiti permettono di ottenere una probabilità di guarigione in base ai parametri noti fino a un dato istante temporale della terapia e consentono inoltre di osservare come potrebbe variare tale probabilità in base alle scelte fatte da quel momento in poi. L'obiettivo finale di questa tesi è proprio quello di proporre 3 modelli predittivi della risposta sostenuta, uno relativo allo stato basale, uno al primo mese e uno al terzo. Grazie alla fase di validazione, tali modelli consentono la costruzione di una terapia mirata rispetto alle varie tipologie di pazienti e dell'adattamento della loro risposta alla terapia in corso di svolgimento. Sebbene il risultato sia relativo alla terapia per l'HCV la base teorica introdotta in questa tesi permette l'applicazione del metodo computazionale per il trattamento di una malattia qualsiasi, o perfino l'applicazione in altri campi.

7.2 Analisi della popolazione in studio

Per lo studio in esame sono stati raccolti 606 casi di pazienti con epatite cronica e cirrosi correlata a infezione da HCV. I dati relativi al loro trattamento sono stati raccolti nel periodo compreso tra Settembre 2005 e Dicembre 2009 presso l'ambulatorio di terapia delle epatiti. Il centro è specializzato nella diagnosi e nella cura di tutte le malattie del fegato, in particolare delle epatiti virali, della cirrosi epatica comprese le sue complicanze e il tumore del fegato (epatocarcinoma).

Dei pazienti reclutati per lo studio, 300 provengono dall'ambulatorio suddetto e 306 provengono da altri centri partecipanti. Per entrambi i gruppi si hanno a disposizione le stesse informazioni relative alle caratteristiche epidemiologiche, cliniche e relative ai regimi antivirali impiegati. Un'analisi statistica svolta a priori ha permesso di verificare che non vi sono state differenze sostanziali nel trattamento. Tali gruppi possono quindi essere utilizzati per un'analisi statistica globale. Come detto nel paragrafo precedente, si sono

individuati i pazienti con genotipo 1 e 4 per analizzarli nel dettaglio, per un totale di 352 soggetti. Le loro caratteristiche vengono descritte in Tabella 7.1.

I pazienti sono stati trattati con Peg-IFN alfa-2a o Peg-IFN alfa-2b in associazione a ribavirina: 121 sono stati trattati con il primo tipo di IFN e 231 con il secondo. L'interferone-pegilato alfa-2b utilizzato è il PegIntron® della Schering-Plough disponibile in dosi di 50, 80, 100, 120 e 150 μg, il Peg-IFN alfa-2a con il quale vengono trattati i pazienti è invece il Pegasys® della Roche il quale è commercializzato in dose standard di 180 μg. Il Peg-IFN alfa-2b permette di essere infatti somministrato in base al peso del soggetto. La scheda tecnica definisce di utilizzare le 5 dosi in commercio rispettivamente per pazienti con peso <40, tra 40 e 64, 65–75,76–85 e >85 kg. L'obiettivo è di trattare il soggetto con la dose pro chilo di 1.5 $\mu g/kg/sett.$ definita dalle linee guida internazionali. La dose consigliata per il Peg-IFN alfa-2a è fissa: 180 $\mu g/kg/sett.$. La ribavirina distribuita come Copegus® (Roche) o Rebetol® (Schering-Plough) è stata impiegata a dosaggio pro chilo al giorno di 15 mg.

I 352 pazienti trattati hanno dimostrato età media 46 anni con una deviazione standard di 11 anni (d'ora in avanti media e deviazione standard vengono indicati con $med \pm d.s$), e hanno seguito la terapia di interferone e ribavirina per un periodo di 37 $\pm$ 16 settimane. L'alta variabilità nella durata è dovuta agli stop prematuri durante il trattamento a causa dell'insorgere di effetti collaterali o della scarsa efficacia conseguita entro il terzo mese (casi NR). I pazienti che hanno assunto Peg-IFN alfa-2a lo hanno fatto per una dose media di 172 $\pm$ 16 μg a settimana, invece i pazienti a cui è stato prescritto Peg-IFN alfa-2b lo hanno assunto con una dose media di 1.21 $\pm$ 0.30 $\mu g/kg$ a settimana. È più corretto descrivere quanti dei pazienti hanno mantenuto la dose standard di 180μg per tutta la terapia piuttosto che calcolare la media per il Peg-IFN alfa-2a, giacché la dose è fissa: essi sono stati 92 su 121 (76.0%) totali che hanno assunto tale interferone. La dose pro chilo giornaliera di RBV media è stata invece di 13.02 $\pm$ 2.07 mg/kg.

Il metodo utilizzato per la determinazione di HCV-RNA virale è stato il COBAS TaqMan HCV della Roche, il quale ha permesso di valutare sia qualitativamente che quantitavamente la viremia di HCV nell'organismo. L'unità di misura standardizzata con cui vengono quantificate le viremie sono le *Unità Internazionali* su millilitro (UI/mL). La sensibilità il range dinamico dello strumento va da 42 UI/mL a 850000 UI/mL con limite minimo di sensibilità a 12.5 UI/mL, per cui è stato assegnato un valore di 10 UI/mL ai casi negativi. La gestione clinica e il monitoraggio della terapia prevedono la valutazione della viremia allo stato basale, a una settimana dalla prima dose, al primo mese dall'inizio della terapia e al terzo mese. Purtroppo questo protocollo non è stato utilizzato per tutti i pazienti a causa della sua recente introduzione nello studio della terapia dell'epatite C: per i pazienti di Padova, si può dire di avere i valori delle viremie basali, del primo mese, del terzo e una parte di quelli relativi alla prima dose di interferone (solo 43), cioè a una settimana dall'inizio del trattamento. Per i pazienti dell'altro gruppo invece non è stata valutata la dose al primo mese nella totalità dei casi: solo per 81 soggetti si può osservare tale valore. Per nessuno dei pazienti del secondo gruppo è inoltre stata misurata la viremia alla prima settimana.

Vi sono studi [1] che indicano che i soggetti che non sopprimono la viremia entro il terzo

mese di trattamento con alta probabilità non otterranno la risposta sostenuta. Perfino quelli che non riescono ad ottenere un abbassamento della viremia di 2 logaritmi in base 10 (che equivale a uno scalo di un fattore 100) dimostrano poche probabilità di eradicare il virus dell'epatite C. Le linee guida internazionali affermano quindi che la terapia per l'HCV può essere sospesa al terzo mese per i suddetti motivi e tale protocollo è stato seguito anche nei centri che hanno seguito i pazienti in esame.

In Figura 7.2 è possibile osservare il grafico con la percentuale cumulativa di pazienti che rispondono alla terapia con negativizzazione dell'HCV-RNA durante il trattamento. Si può

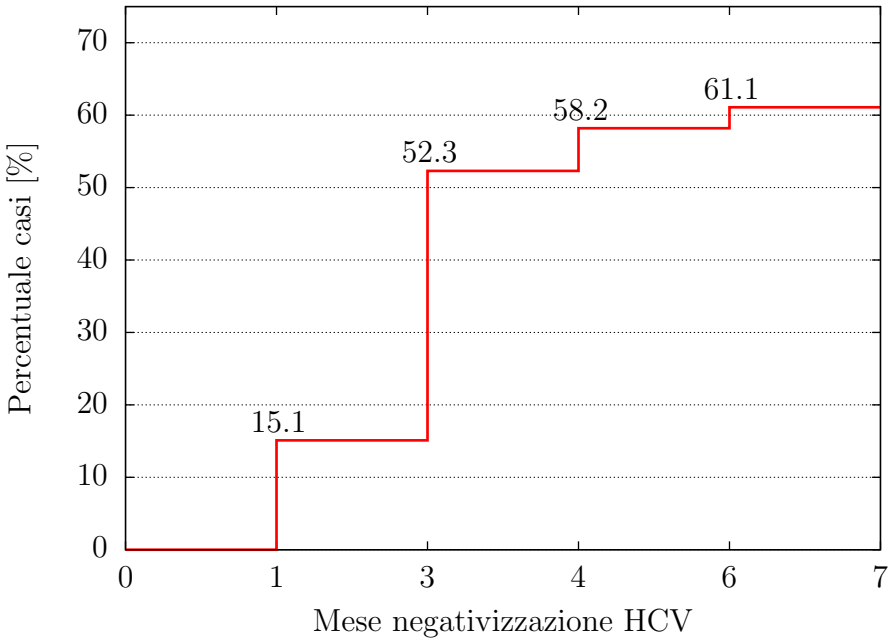


Figura 7.2: Percentuale di pazienti che rispondono durante il trattamento

notare che le percentuali di negativi al test dell'HCV sono trascurabili una volta superato il terzo mese con viremia ancora positiva: questo è dovuto anche alla prematura sospensione del trattamento da linee guida. La distribuzione delle risposte dei pazienti durante la terapia per il campione in esame è rappresentata nel diagramma a torta di Figura 7.3. La nomenclatura è relativa alla Tabella 1.1 del Capitolo 1 e viene qui riproposta: se il paziente è negativo al primo mese si definisce una *Rapid Virological Response* (RVR), se al terzo una *complete Early Virological Response* (cEVR) e, nel caso tardivo, cioè tra il terzo e il sesto mese di trattamento, di una *partial Early Virological Response* (pEVR). Si noti come la probabilità di non ottenere la soppressione della viremia durante la terapia, quindi di essere un "Not Responder" (NR), sia stata del 38.9%. È logico che la probabilità di non ottenere la risposta a lungo termine del 55.4% (in Figura 7.1) sia dovuta anche ai relapser (RR). Tali casi sono il 16.5% del totale dei pazienti. Da ora in poi un paziente che non ottiene la risposta sostenuta verrà identificato come "Non LTR".

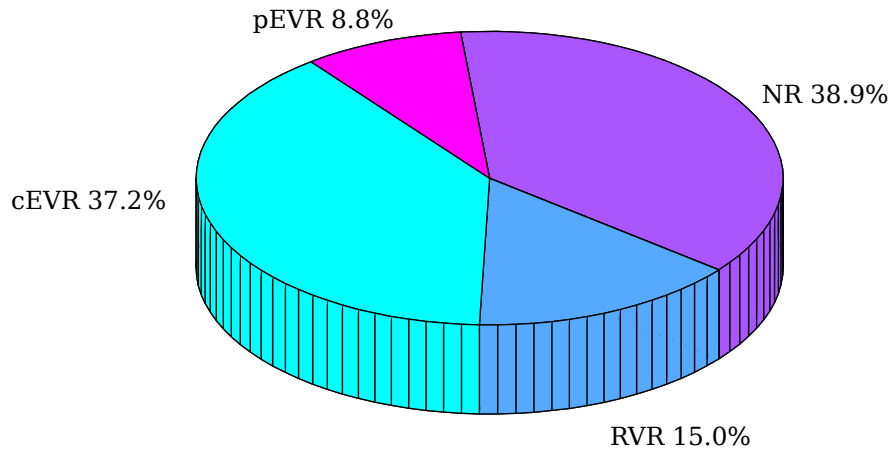


Figura 7.3: Distribuzione dei vari profili di risposta alla terapia antivirale nei pazienti con HCV genotipo 1 e 4

La Figura 7.4 mostra invece come la più precoce capacità di soppressione della viremia condizioni l'efficacia della terapia: tra coloro che negativizzano l'HCV-RNA già al primo mese, ben il 90.6% otterrà una risposta sostenuta; tale probabilità andrà a ridursi fino al valore minimo del 30% per coloro con test negativo solo al sesto mese. Osservando tale grafico si può notare che le probabilità di guarire completamente dall'epatite C diminuiscono per chi non è in grado di sopprimere la carica virale in breve tempo. Questo fatto valida la decisione di costruire i modelli relativamente ai soli istanti temporali citati nell'introduzione di questo capitolo. Un modello per lo stato basale permette di prendere decisioni su tutta la terapia che si sta per intraprendere e quelli al primo mese e al terzo permettono di valutare l'andamento dinamico di tale terapia. Non vengono costruiti modelli al quarto, quinto o sesto mese in quanto non particolarmente utili alla gestione clinica del trattamento per i motivi suddetti.

7.2.1 Studio dei parametri

Nello studio proposto vengono considerati diversi parametri *clinici* che potrebbero essere utili per conoscere la possibilità di risposta trattamento. Tale analsi è stata utile per capire meglio quali sono le variabili che identificano meglio la risposta a lungo termine. Questo paragrafo ha quindi come scopo principale la presentazione dei dati studiati per avere una visione d'insieme del problema.

Tutte le variabili sono riportate in Tabella 7.1. Essa è suddivisa in base ai parametri relativi alla *terapia* ed al *paziente*. Oltre alle informazioni quali peso, altezza e IMC (Indice Massa Corporea), nella parte relativa al paziente è possibile trovare parametri bioumorali e istologici. Le variabili categoriche sono tutte binarie, per esse viene indicata la frequenza relativa dei loro valori nel campione di 352 soggetti. Per le variabili continue si presenta il valore medio e la sua deviazione standard, nonché il loro valore minimo e massimo.

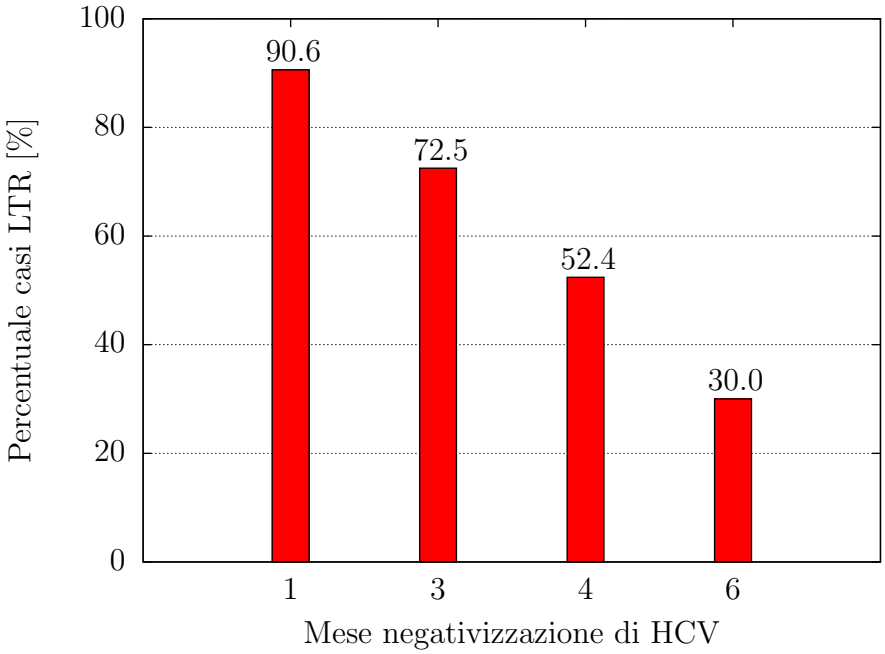


Figura 7.4: Percentuale di pazienti che ottengono la risposta sostenuta in base al mese di negativizzazione di HCV

Nello specifico, vengono studiati i parametri istologici, relativi al grado di progressione della malattia, grading e staging secondo la classificazione di Ishak [33]. Entrambi vengono valutati dal medico tramite osservazione di tessuto epatico prelavato da biopsia. Il *grading* è una valutazione istologica dell'attività necroinfiammatoria che quantifica quindi il grado di flogosi/infiammazione del fegato, esso va da un minimo di 0 a un massimo di 18. Lo *staging* è invece una valutazione istologica del grado di fibrosi del fegato e quantifica il livello di progressione della malattia. Il suo minimo è 0 e il massimo viene raggiunto nel caso di cirrosi avanzata, tale stadio viene identificato con il valore 6 di staging. Altro parametro istologico studiato è il grado di steatosi epatica o percentuale di grassi accumulati nel fegato. Un valore alto di steatosi indica un malfunzionamento del metabolismo epatico. Bisogna far notare che i test istologici, derivati da una biopsia, che è un esame altamente invasivo, si riferiscono allo stato basale del soggetto, prima dell'inizio della terapia.

I parametri clinici relativi ad istanti temporali diversi vengono ottenuti allo stato basale (prima dell'inizio della terapia) e durante la terapia al primo e terzo mese. Solo la viremia in alcuni dei casi è stata valutata anche a una settimana dall'inizio del trattamento. I valori bioumorali studiati sono stati l'*emocromo*, valutando la quantità di la quantità di emoglobina, i globuli bianchi e le piastrine; il livello di transaminasi: la ALT (Alanino amino transferasi), che è uno dei principali indicatori di danno al fegato, la AST (Aspartato transferasi) e la GGT o γGT (Gamma glutamil transpeptidasi). Pur essendo esami di routine in alcuni casi tali test non vengono effettuati a tutti gli istanti temporali citati (basale, primo mese, terzo mese), perciò la Tabella 7.1 evidenzia la percentuale di dati

mancanti.

Si vogliono inoltre presentare le differenze tra il gruppo di pazienti che hanno ottenuto la risposta sostenuta (LTR) e quelli che non sono riusciti a farlo (non LTR). I test statitici per confrontare i due insiemi di soggetti sono il test del *Chi-quadro* per le variabili categoriche binarie e il test di *Student* per le variabili continue. Tali test sono stati trattati nel dettaglio nel Capitolo 2. Si è scelto di fissare il livello di significatività α a 0.05, vengono pertanto considerati significativi valori di p-value inferiori o uguali a tale soglia.

I confronti visibili in Tabella 7.2 sono stati suddivisi rispetto alle variabili relative alla terapia e quelle relative al paziente come fatto nella precedente tabella (Tabella 7.1). Per coerenza si è mantenuto lo stesso ordine delle variabili rispetto ad essa. È stato però aggiunto il confronto tra la frequenza di pazienti che mantengono la dose costante a 180 μg a settimana per il Peg-IFN alfa-2a negli LTR e nei non LTR.

Si noti anche come la cirrosi sia un fattore sfavorevole per il conseguimento dell'efficacia terapeutica. L'identificazione di un paziente cirrotico o con epatite cronica viene fatta in base a test clicini e istologici, tra cui la valutazione dello staging. Oltre ai valori di viremia più elevati per i non LTR rispetto agli LTR, sembrano essere significativi i valori delle transaminasi per discriminare gli LTR dagli altri. Questo risultato è plausibile ed associabile al fatto che l'alterazione dei valori di ALT è correlata con la citolisi epatica dovuta al danno epatico causato dal virus specialmente se resistente alla terapia.

Generalmente i pazienti LTR sono risultati un po' più giovani degli altri ed i casi in sovrappeso hanno ottenuto una percentuale minore di risposte.

È doveroso sottolineare che i parametri in esame sono un sottoinsieme di tutti quelli disponibili per la possibile analisi. La selezione è avvenuta in base al parere tecnico di medici epatologi che hanno identificato quelli più significativi per predire la risposta sostenuta. La scelta ha tenuto conto anche della facilità nel reperire un determinato attributo. Un metodo efficace di predizione deve basarsi su parametri semplici e facilmente misurabili per poter essere effettivamente utilizzato nella pratica. Per questo motivo i modelli costruiti nei capitoli successivi *non* utilizzano né il grading né lo staging né il grado di steatosi di un paziente. Tali misure vengono fatte valutando un campione di tessuto epatico prelevato da biopsia, che è un esame altamente invasivo ed ormai superato da altre tecniche non invasive.

7.3 Approccio risolutivo

Questa tesi è incentrata sia sull'attività di *problem solving* che su quella di *problem posing*. Per quanto il problema di predizione della risposta sostenuta sia ben definito, il problema generale è estrapolare più informazioni possibili dai dati a disposizione per curare i pazienti. È stato infatti necessario in corso d'opera definire quali erano gli aspetti significativi del problema per poterlo meglio caratterizzare.

In principio ci si è concentrati sul problema della pura predizione. Il fenomeno in esame (la guarigione del paziente) è stato inizialmente studiato nella condizione di semplici

osservatori. Sotto tale ipotesi l'unica possibilità è quella di analizzare l'andamento dei parametri clinici del paziente e inferire sulla probabilità di riuscita del trattamento.

La prima tecnica utilizzata per costruire dei modelli è stata l'induzione di alberi decisionali o *decision tree* (si veda il Capitolo 5). L'obiettivo è stato quello di costruire tre alberi decisionali: uno allo stato basale del paziente, uno al primo mese e uno al terzo. I modelli così generati sono utili per discriminare i nuovi pazienti tra LTR e non LTR. Un albero all'istante basale permette di identificare come possibile long term responder o meno un soggetto in base alle sue caratteristiche prima del trattamento. Gli alberi costruiti invece al primo e terzo mese hanno lo stesso obiettivo e posseggono, oltre alle informazioni basali del paziente, quelle relative alla finestra temporale rispettivamente prima del primo mese e prima del terzo. La predizione tra LTR e non LTR è ovviamente più accurata più la terapia avanza grazie al fatto che i modelli sono costruiti rispetto a quantità di dati via via maggiori.

Un albero decisionale relativo allo stato basale prodotto utilizzando 352 pazienti in studio è presentato in Figura 7.5. Le caselle quadrate, che rappresentano le foglie, sono

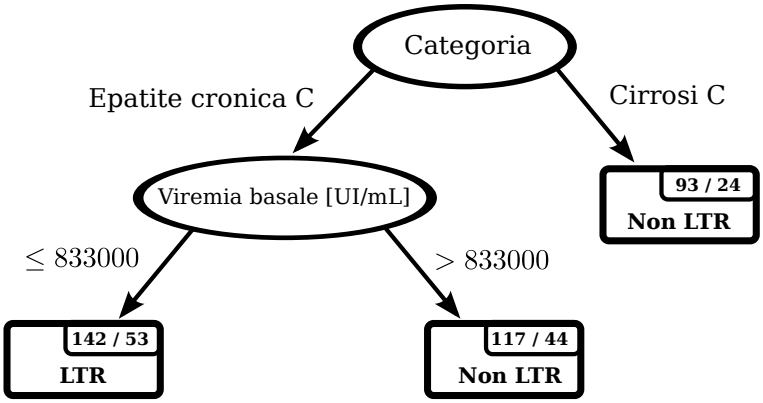


Figura 7.5: Decision tree indotto con i 352 pazienti studiati

etichettate con "LTR" o "non LTR" a seconda della classe di maggioranza dei pazienti per tale nodo. Le foglie inoltre sono segnate in alto a destra con la dicitura X/Y dove X è il numero di pazienti che finiscono in quel nodo e Y il numero pazienti erroneamente classificati da quel nodo. Si prenda in considerazione il nodo più profondo e a sinistra etichettato con LTR. In quel nodo finiscono i pazienti con epatite cronica e viremia basale minore o uguale a 833000 UI/mL. Tali pazienti sono 142, ma ben 53 di essi in realtà non sono LTR. Un albero del genere è utile per poter classificare automaticamente un nuovo paziente prima del trattamento. Osservando solamente due parametri, quali la viremia basale e la sua categoria (ottenibile tramite test clinici o istologici), egli viene identificato a priori come LTR o meno. La scelta dei parametri non è banale ed è descritta nel Capitolo 5. Il metodo computazionale che produce la suddivisione ai nodi sceglie il parametro e il corrispettivo test per suddividere al meglio i pazienti LTR dai non LTR. In base al supporto di un modello del genere il medico è in grado di fare assunzioni su quale sia la terapia più adatta per il soggetto.

A questo punto dell'analisi ci si è accorti che è molto difficile stabilire a priori se il paziente sarà LTR o meno, in quanto la distinzione tra queste due classi non è netta. I parametri in studio, a causa dell'alta variabilità della risposta dei soggetti, non permettono una classificazione del nuovo paziente in "LTR" o "non LTR" in maniera accurata. Si è scelto allora l'approccio probabilistico dei PET (*Probability Estimation Tree*), trattati nel Capitolo 6. Il risultato della previsione, da quel momento in poi, non è stata più una classe ma la probabilità di appartenere a una data classe. Nel caso in esame la probabilità è quella di essere un LTR.

È noto dalla teoria dei PET che le migliori stime di probabilità vengono fatte quando l'albero va più in profondità rispetto a un decision tree classico. Pertanto i modelli prodotti sono leggermente più corposi ma non più complessi. Uno scorcio di PET indotto con i pazienti dello studio è mostrato in Figura 7.6. Si noti come sia chiaro osservare quali sono i parametri che più apportano cambiamento alla probabilità di risposta. Ogni nodo mostra la percentuale di pazienti che rispondono alla sua destra, si ricordi però che solo i nodi terminali vengono utilizzati per la stima di probabilità di un nuovo paziente. Tutte le misure di valutazione di un PET infatti si riferiscono alle stime di probabilità indotte dalle foglie, la presenza della stima ai nodi è solo un'informazione locale. Ogni nodo, come

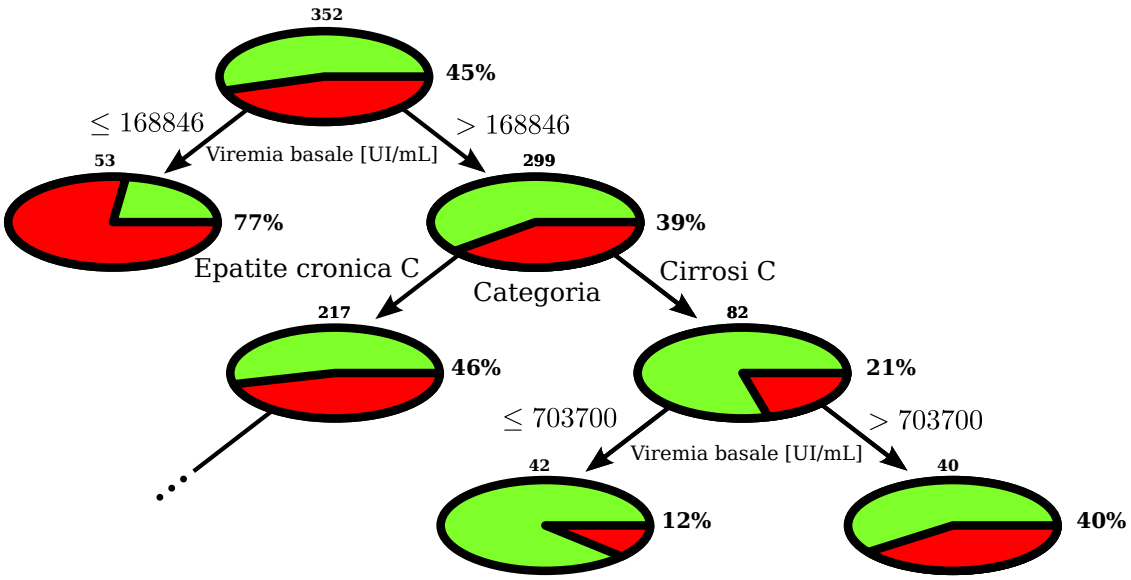


Figura 7.6: Probability Estimation Tree (PET) indotto con i 352 pazienti studiati

nei decision tree, presenta il numero di pazienti a lui associati.

Osservando l'ultimo nodo interno (che esegue un test sulla viremia basale) si nota come la probabilità di rispondere sia del 12% per i pazienti cirrotici con viremia basale nell'intervallo $[168846, 703700]$ e del 30% per i pazienti cirrotici con viremia al di sopra delle 703700 UI/mL. Tale suddivisione *non* sarebbe mai stata indotta da un decision tree in quanto porterebbe ad etichettare entrambe le foglie con "Non LTR".

I modelli appena introdotti fanno assunzioni implicite sulla terapia: essi ipotizzano che tutti i pazienti siano stati trattati in modo identico. Tale assunzione è naturalmente

sbagliata. Non solo, spesso un paziente è costretto a scalare (diminuire) la dose di farmaco a causa di complicanze o effetti collaterali. Quindi la stessa terapia nel singolo paziente *è variabile*. Il tentativo di introdurre le dosi di farmaco come fattore predittivo della risposta sostenuta ha evidenziato la vera natura del problema. Non basta mettersi nella condizione di osservatori del fenomeno in quanto questa è una vera e propria approssimazione. Il processo di guarigione è direttamente indotto dai farmaci utilizzati e la modifica delle loro dosi apporta cambiamenti sostanziali alla risposta del paziente alla terapia.

La Figura 7.7 serve per modellare il nuovo problema. In questa figura viene mostrato

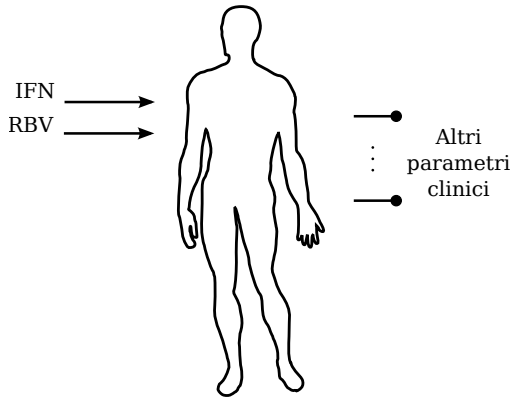


Figura 7.7: I farmaci vengono forniti al paziente e gli altri attributi possono essere utilizzati per valutare l'andamento della terapia

graficamente quello che si intende realizzare. L'interferone e la ribavirina non possono essere usati alla stregua degli altri parametri. Sono *variabili indipendenti* che modificate apportano cambiamenti al processo di guarigione. La Figura 7.7 evidenzia come l'*IFN* e la *RBV* vengano forniti al paziente, e quindi possano essere modificati dal medico, e come gli *altri parametri clinici* possano essere utilizzati per valutare l'andamento della terapia.

Si è allora voluto costruire un modello che fornisca non solo la predizione della probabilità di risposta a lungo termine per un nuovo paziente ma che permetta di ipotizzare come possa variare in base alle *scelte* che si compiono. Per ottenerlo è necessario separare i parametri che si possono modificare per ottenere esisti diversi del processo da quelle che si possono solo osservare. Le variabili che fungono da "manopole" per ottenere esiti diversi di terapia sono: la dose di IFN, la dose di RBV, la durata del trattamento di IFN e la durata del trattamento di RBV. Un modello che tiene conto di quanto appena detto è formalizzato nel Capitolo 8 e presentato nel paragrafo successivo.

7.4 PET incrementale

Alla fine del paragrafo precedente si è parlato di quali sono le variabili su cui si può agire per modificare l'esito della terapia. Dato un istante temporale t della terapia di un nuovo paziente è però possibile modificare l'evoluzione della terapia stessa solo da quell'istante in poi. Parametri relativi a istanti temporali $> t$ vengono chiamati *futuri e modificabili*.

Gli attributi osservabili studiando l'andamento della terapia dei pazienti nella finestra temporale $[0, t]$ vengono invece chiamati *passati e fissi*.

Il nuovo modello proposto è un PET che permette di stimare la probabilità di risposta sostenuta alla terapia all'istante temporale generico t in modo deterministico osservando i valori degli attributi passati e fissi. A un nuovo paziente può così essere assegnata una probabilità in base ad informazioni note in quanto relative al periordo $[0, t)$, così come veniva fatto dai PET del paragrafo precedente. In più ogni nodo foglia, se possibile, viene ulteriormente suddiviso in base ai parametri che possono essere modificati nel futuro. In questo modo si separa logicamente il passato dal futuro. Un'eventuale scissione di un nodo foglia fornisce un supporto per le decisioni future del medico. Tale albero viene chiamato *PET incrementale*.

I parametri utilizzati in questa tesi sono rappresentati in Tabella 7.3. In questo lavoro ci si è voluti concentrare sulle sole dosi di interferone e ribavirina. L'obiettivo è quindi ottenere un PET incrementale che permetta allo stato basale di assegnare a un nuovo paziente una probabilità rispetto a semplici test clinici e poi consigli qual'è la dose più indicata per iniziare il trattamento. Poi altri due PET incrementali, uno al primo mese e uno al terzo, che osservino l'andamento dei parametri clinici e la quantità di dose assunta fino all'istante temporale in questione e assegnano la probabilità di essere LTR in modo classico rispetto a tali parametri passati e fissi. Entrambi offrono poi, come nel caso dell'albero costruito a basale, un consiglio su come procedere la terapia, rispettivamente dal primo mese e dal terzo mese. I parametri clinici citati in Tabella 7.3 sono quelli descritti nello studio dei 352 pazienti in esame (Paragrafo 7.2.1). Come già indicato vengono eliminati il grading, lo staging e il grado di steatosi perché ridondanti rispetto all'attributo "Categoria". Sempre per motivi di ridondanza non è stato fornito in input al metodo di induzione nemmeno l'attributo "Altezza" in modo da obbligare split su "Peso" o "IMC".

I rapporti delle dosi vengono calcolati indipendentemente uno dall'altro ma con lo stesso metodo. Qui viene spiegato come ottenere il valore dell'attributo "R. IFN [%]" per un paziente che ha completato la terapia. L'unica accortezza che bisogna avere è di utilizzare la dose pro chilo per il Peg-IFN alfa-2b. In genere ai pazienti trattati per l'epatite C viene concesso al massimo un solo scalo per non precludere le possibilità di guarigione. Se la terapia non sortisce gli effetti desiderati, si preferisce terminare prematuramente il trattamento piuttosto che scalare ulteriormente la dose. Si supponga che il paziente assuma la dose d_1 settimanale di IFN all'inizio della terapia e che la mantenga per un tempo t_1. Si supponga poi che per qualche motivo essa venga scalata alla dose d_2, la quale viene mantenuta per un tempo t_2. Per quanto detto sopra, $t_1 + t_2$ è la durata effettiva della terapia del paziente in esame. La dose media di interferone assunto si calcola come:

$$\frac{d_1 \cdot t_1 + d_2 \cdot t_2}{t_1 + t_2}$$

Il rapporto è ottenuto dividendo la dose media per la dose standard, cioè 180 $\mu g/sett.$ per il Peg-IFN alfa-2a e 1.5 $\mu g/kg/sett.$ per il Peg-IFN alfa-2b. Il rapporto di IFN di un paziente già trattato può essere utilizzato per costruire il modello allo stato basale per un nuovo paziente. Tale parametro viene calcolato in generale per un qualsiasi istante t. La

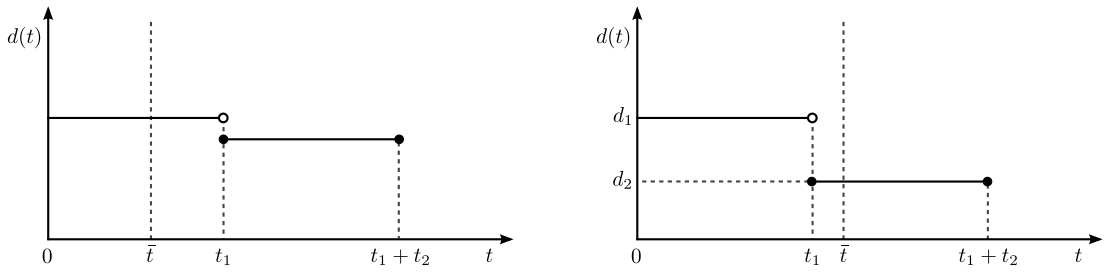
formula di seguito introdotta serve per calcolare il valore dei rapporti al primo e al terzo mese. La condizione per considerare un soggetto nella costruzione di un PET per predire la risposta di un nuovo paziente all'istante t è che $t_1 + t_2 > t$. Semplicemente, il soggetto deve avere una durata di terapia *superiore* all'istante in cui si vuole andare a valutare il rapporto: altrimenti non può essere considerato nell'analisi. Questo problema viene risolto lasciando il campo relativo a tale paziente *vuoto*. È possibile calcolare la dose media di IFN assunto *prima* dell'istante t con le seguenti formule: se $t < t_1$ essa è d_1 e se $t \geq t_1$ la dose media è

$$\frac{d_1 \cdot t_1 + d_2 \cdot (t - t_1)}{t}$$

Analogamente la dose media di IFN assunta *dopo* l'istante t è così calcolata: se $t < t_1$ essa è

$$\frac{d_1 \cdot (t_1 - t) + d_2 \cdot t_2}{t_1 + t_2 - t}$$

e se $t \geq t_1$ essa è d_2 Per rendere più chiaro l'utilizzo di queste formule sono stati rappresentati due andamenti temporali di dosi in Figura 7.8. Si pensi di voler calcolare i rapporti a $\bar{t}$. Un caso ha $\bar{t} < t_1$ e l'altro $\bar{t} \geq t_1$.



(a) $\bar{t} < t_1$ (b) $\bar{t} \geq t_1$

Figura 7.8: Variazione della dose rispetto al tempo

Le dosi medie vengono poi divise per la dose standard prevista dalla terapia. Questa operazione permette di utilizzare i pazienti trattati con interferone-pegilato alfa-2a e alfa-2b indistintamente e rendendo quindi inutile l'attributo "N. Trattati Peg-IFN" li discriminava. Entrambi avranno infatti un attributo comune relativo al rapporto *adimensionale*. Si ricorda che il procedimento per calcolare i rapporti è valido anche per la RBV, dove viene utilizzata una dose standard pro chilo di 15 $mg/kg/die$.

Split sugli attributi relativi ai rapporti di IFN e RBV futuri forniscono un'informazione necessaria per la terapia di un nuovo paziente. Essi presentano come sia possibile variare la terapia da quel momento in poi e che risultati si otterrebbero. L'utilizzo dei rapporti in base alle sole dosi di IFN e RBV è il primo approccio al problema sviluppato in questa tesi. È chiaramente possibile adattare la tecnica per inferire anche su quello che dovrebbe essere la durata di terapia più adeguata per un qualsiasi soggetto.

Un esempio di PET incrementale ottenuto per stimare la probabilità di essere LTR allo stato basale è presentato in Figura 7.9. Questo è stato ottenuto dai dati dei 352 pazienti

in esame e non è rappresentato completamente in figura per evidenziare gli aspetti più importanti. Si ipotizzi di avere un soggetto con viremia basale di 100000 UI/mL. In modo

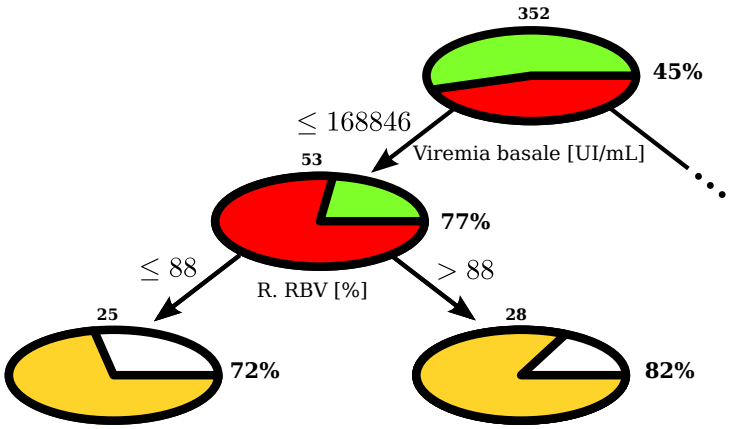


Figura 7.9: Esempio di PET incrementale

semplice è possibile stimare la sua probabilità di guarigione percorrendo i nodi relativi al passato nell'albero in figura. Tale probabilità è del 77%. Ma cosa è possibile far per migliorare il suo trattamento? Il suggerimento del modello è che il paziente non vada sotto l'89% della dose standard di ribavirina mediamente assunta durante il suo trattamento. Nel dettaglio questo specifica di non andare sotto i 13.2 mg/kg giornalieri. Se non si segue l'indicazione del modello è possibile che il paziente risponda alla terapia con più difficoltà. Le probabilità stimate di essere LTR sono 82% e 72% per dosi rispettivamente superiori e inferiori o uguali all'88% della dose standard pro chilo di RBV.

Tale informazione è realmente di utile supporto alle decisioni del medico. La necessità di uno scalo della dose di IFN o RBV, nonché le frequenti richieste di alleviare gli effetti collaterali da parte dei pazienti, possono essere valutate con i modelli ottenuti, in modo da avere una migliore prospettiva sulle conseguenze di determinate scelte. Inoltre, la decisione di sospensione precoce del trattamento a causa della sua poca efficacia potrebbe essere supportata. Nel caso che l'albero non suggerisca nessuna informazione sul futuro del trattamento è comunque possibile prendere in considerazione la parte costruita rispetto ai dati passati per stimare le probabilità di ottenere la risposta sostenuta.

(a) Relativi al trattamento

Attributo	(%)	**Caratteristiche**
Dose Peg-IFN alfa-2a [$\mu g/sett.$]	0.0	(med. ± d.s.) 172 ± 16 (min) 108 (max) 180
Dose Peg-IFN alfa-2b [$\mu g/kg/sett.$]	0.0	(med. ± d.s.) 1.21 ± 0.30 (min) 0.52 (max) 2.03
Dose RBV [$mg/kg/die$]	0.0	(med. ± d.s.) 13.02 ± 2.07 (min) 7.59 (max) 18.18
Durata trattamento [$settimane$]	0.0	(med. ± d.s.) 37 ± 16 (min) 4 (max) 72
N. Trattati Peg-IFN [alfa-2a/alfa-2b]	0.0	(%) 34.4/65.6
Trattamento [naive/ritrattatamento]	0.0	(%) 64.2/35.8

(b) Relativi al paziente

Attributo	(%)	**Caratteristiche**
Sesso [Maschio/Femmina]	0.0	(%) 69.9/30.1
Genotipo [HCV1/ HCV4]	0.0	(%) 90.6/9.4
Categoria [Epatite cronica C/Cirrosi C]	0.0	(%) 73.6/26.4
Durata malattia [$mesi$]	8.5	(med. ± d.s.) 198 ± 126 (min) 6 (max) 510
Peso [kg]	0.0	(med. ± d.s.) 74 ± 13 (min) 46 (max) 129
Altezza [cm]	3.1	(med. ± d.s.) 171.79 ± 8.68 (min) 150.00 (max) 195.00
IMC [kg/m^2]	3.1	(med. ± d.s.) 25.05 ± 3.56 (min) 17.36 (max) 39.56
Età [$anni$]	0.0	(med. ± d.s.) 46 ± 11 (min) 21 (max) 74
Viremia basale [UI/mL]	0.0	(med. ± d.s.) 566063 ± 303810 (min) 4650 (max) 850000
Viremia 1^a settimana [UI/mL]	87.8	(med. ± d.s.) 232204 ± 317468 (min) 40 (max) 850000
Viremia 1° mese [UI/mL]	40.3	(med. ± d.s.) 101236 ± 204883 (min) 10 (max) 850000
Viremia 3° mese [UI/mL]	0.6	(med. ± d.s.) 59447 ± 147234 (min) 10 (max) 850000
Emoglobina basale [g/L]	5.1	(med. ± d.s.) 147.81 ± 19.38 (min) 13.40 (max) 180.00
Globuli Bianchi basale [$n. \cdot 10^9/L$]	5.1	(med. ± d.s.) 6.43 ± 1.72 (min) 2.60 (max) 15.20
Piastrine basale [$n. \cdot 10^9/L$]	5.1	(med. ± d.s.) 211.10 ± 64.56 (min) 0.01 (max) 426.00
AST basale [U/L]	17.6	(med. ± d.s.) 79.34 ± 52.81 (min) 19.30 (max) 400.50
ALT basale [U/L]	4.8	(med. ± d.s.) 148.95 ± 118.95 (min) 19.60 (max) 1106.90
GGT basale [U/L]	39.8	(med. ± d.s.) 92.44 ± 77.20 (min) 9.60 (max) 443.30
Emoglobina 1° mese [g/L]	3.4	(med. ± d.s.) 128.09 ± 19.73 (min) 9.70 (max) 168.00
Globuli bianchi 1° mese [$n. \cdot 10^9/L$]	3.4	(med. ± d.s.) 4.06 v 1.47 (min) 1.02 (max) 11.90
Piastrine 1° mese [$n. \cdot 10^9/L$]	4.5	(med. ± d.s.) 172.85 ± 66.39 (min) 39.00 (max) 796.00
AST 1° mese [U/L]	19.3	(med. ± d.s.) 50.75 ± 34.18 (min) 14.00 (max) 344.30
ALT 1° mese [U/L]	5.4	(med. ± d.s.) 76.48 ± 62.96 (min) 14.80 (max) 579.50
GGT 1° mese [U/L]	71.9	(med. ± d.s.) 79.65 ± 69.92 (min) 9.00 (max) 425.50
Emoglobina 3° mese [g/L]	6.0	(med. ± d.s.) 122.15 ± 18.17 (min) 10.00 (max) 165.00
Globuli bianchi 3° mese [$n. \cdot 10^9/L$]	6.0	(med. ± d.s.) 3.48 ± 1.18 (min) 1.44 (max) 9.03
Piastrine 3° mese [$n. \cdot 10^9/L$]	7.4	(med. ± d.s.) 156.47 ± 56.39 (min) 47.00 (max) 377.00
AST 3° mese [U/L]	20.7	(med. ± d.s.) 47.23 ± 33.16 (min) 15.00 (max) 232.30
ALT 3° mese [U/L]	8.5	(med. ± d.s.) 61.96 ± 59.07 (min) 12.40 (max) 514.30
GGT 3° mese [U/L]	70.7	(med. ± d.s.) 70.41 ± 70.07 (min) 14.80 (max) 482.63
Steatosi [$grado$]	10.8	(med. ± d.s.) 1 ± 1 (min) 0 (max) 3
Grading [$Ishak$]	13.1	(med. ± d.s.) 6 ± 2 (min) 2 (max) 18
Staging [$Ishak$]	13.1	(med. ± d.s.) 3 ± 1 (min) 1 (max) 6

Tabella 7.1: Presentazione delle variabili studiate, la seconda colonna indica la percentuale di valori mancanti

(a) Relativi al trattamento

	LTR	**Non LTR**	p-value
Dose Peg-IFN alfa-2a [$\mu g/sett.$] (media ± s.d.)	174 ± 13	171 ± 18	0.323
Dose Peg-IFN alfa-2a [standard/scalata] (%)	81.4/18.6	71.0/29.0	0.181
Dose Peg-IFN alfa-2b [$\mu g/kg/sett.$] (media ± s.d.)	1.26 ± 0.26	1.18 ± 0.33	≤ 0.05
Dose RBV [$mg/kg/die$] (media ± s.d.)	13.29 ± 1.93	12.80 ± 2.15	≤ 0.05
Durata trattamento [$settimane$] (media ± s.d.)	46 ± 11	30 ± 16	≤ 0.05
N. Trattati Peg-IFN [alfa-2a/alfa-2b] (%)	37.6/62.4	31.8/68.2	0.256
Trattamento [naive/ritrattatamento] (%)	66.2/33.8	62.6/37.4	0.474

(b) Relativi al paziente

	LTR	**Non LTR**	p-value
Sesso [Maschio/Femmina] (%)	72.0/28.8	68.2/31.8	0.444
Genotipo [HCV1/HCV4] (%)	89.2/10.8	91.8/8.2	0.401
Categoria [Epatite cronica C/Cirrosi C] (%)	84.7/15.3	64.6/35.4	≤ 0.05
Durata malattia [$settimane$] (media ± s.d.)	182 ± 115	211 ± 133	≤ 0.05
Peso [kg] (media ± s.d.)	74 ± 14	74 ± 13	0.693
Altezza [cm] (media ± s.d.)	173 ± 8	171 ± 9	0.070
IMC [kg/m^2] (media ± s.d.)	24.67 ± 3.55	25.35 ± 3.55	0.080
Età [$anni$] (media ± s.d.)	44 ± 10	48 ± 11	≤ 0.05
Viremia basale [UI/mL] (media ± s.d.)	499428 ± 325629	619713 ± 274311	≤ 0.05
Viremia 1^a settimana [UI/mL] (media ± s.d.)	23085 ± 35051	368936 ± 345477	≤ 0.05
Viremia 1° mese [UI/mL] (media ± s.d.)	5879 ± 23825	183090 ± 251417	≤ 0.05
Viremia 3° mese [UI/mL] (media ± s.d.)	38 ± 186	107775 ± 184852	≤ 0.05
Emoglobina basale [g/L] (media ± s.d.)	146.31 ± 22.64	148.99 ± 16.33	0.209
Globuli Bianchi basale [$n. \cdot 10^9/L$] (media ± s.d.)	6.60 ± 1.76	6.30 ± 1.68	0.115
Piastrine basale [$n. \cdot 10^9/L$] (media ± s.d.)	221.88 ± 55.37	202.52 ± 70.00	≤ 0.05
AST basale [U/L] (media ± s.d.)	74.32 ± 58.30	83.60 ± 47.43	0.136
ALT basale [U/L] (media ± s.d.)	151.16 ± 133.46	147.16 ± 106.08	0.760
GGT basale [U/L] (media ± s.d.)	65.64 ± 49.98	116.38 ± 88.74	≤ 0.05
Emoglobina 1° mese [g/L] (media ± s.d.)	129.03 ± 16.98	127.35 ± 21.71	0.436
Globuli Bianchi 1° mese [$n. \cdot 10^9/L$] (media ± s.d.)	4.22 ± 1.47	3.94 ± 1.46	0.078
Piastrine 1° mese [$n. \cdot 10^9/L$] (media ± s.d.)	180.32 ± 52.66	166.97 ± 75.08	0.067
AST 1° mese [U/L] (media ± s.d.)	40.99 ± 31.96	59.22 ± 33.88	≤ 0.05
ALT 1° mese [U/L] (media ± s.d.)	58.07 ± 56.11	91.22 ± 64.39	≤ 0.05
GGT 1° mese [U/L] (media ± s.d.)	57.53 ± 60.94	98.85 ± 72.05	≤ 0.05
Emoglobina 3° mese [g/L] (media ± s.d.)	121.97 ± 17.14	122.29 ± 19.04	0.0.875
Globuli Bianchi 3° mese [$n. \cdot 10^9/L$](media ± s.d.)	3.53 ± 1.16	3.43 ± 1.20	0.432
Piastrine 3° mese [$n. \cdot 10^9/L$](media ± s.d.)	164.41 ± 49.10	149.87 ± 61.16	≤ 0.05
AST 3° mese [U/L] (media ± s.d.)	37.18 ± 24.41	55.75 ± 37.07	≤ 0.05
ALT 3° mese [U/L] (media ± s.d.)	43.82 ± 35.60	77.01 ± 69.62	≤ 0.05
GGT 3° mese [U/L] (media ± s.d.)	51.77 ± 72.35	82.73 ± 66.26	≤ 0.05
Steatosi [$grado$] (media ± s.d.)	0.78 ± 0.77	1.01 ± 0.78	≤ 0.05
Grading [$Ishak$] (media ± s.d.)	5 ± 2	6 ± 2	≤ 0.05
Staging [$Ishak$] (media ± s.d.)	3 ± 1	3 ± 1	≤ 0.05

Tabella 7.2: Differenza tra pazienti che ottengono la risposta sostenuta e quelli che non la ottengono

Basali	1° mese terapia	3° mese terapia
	Passati e Fissi	
∗ Parametri clinici basali	∗ Parametri clinici basali + 1° mese ∗ Rapporto IFN assunto rispetto alla dose standard fino al 1° mese **(R. IFN prima 1°m.**[%]**)** ∗ Rapporto RBV assunta rispetto alla dose standard fino al 1° mese **(R. RBV prima 1°m.**[%]**)**	∗ Parametri clinici basali + 1° mese + 3° mese ∗ Rapporto IFN assunto rispetto alla dose standard fino al 3° mese **(R. IFN prima 3°m.**[%]**)** ∗ Rapporto RBV assunta rispetto alla dose standard fino al 3° mese **(R. RBV prima 3°m.**[%]**)**
	Futuri e Manipolabili	
∗ Rapporto IFN assunto rispetto alla dose standard **(R. IFN** [%]**)** ∗ Rapporto RBV assunta rispetto alla dose standard **(R. RBV** [%]**)**	∗ Rapporto IFN assunto rispetto alla dose standard dopo il 1° mese **(R. IFN dopo 1°m.**[%]**)** ∗ Rapporto RBV assunta rispetto alla dose standard dopo il 1° mese **(R. RBV dopo 1°m.**[%]**)**	∗ Rapporto IFN assunto rispetto alla dose standard dopo il 3° mese **(R. IFN dopo 3°m.**[%]**)** ∗ Rapporto RBV assunta rispetto alla dose standard dopo il 3° mese **(R. RBV dopo 3°m.**[%]**)**

Tabella 7.3: Parametri utilizzati nel modello a supporto delle decisioni

Capitolo 8

Analisi del problema informatico

In questo capitolo viene riproposto il procedimento logico fatto nel capitolo precedente per arrivare alla soluzione adottata. Tutti i risultati vengono presentati cercando di andare nel dettaglio matematico. L'unica differenza è che ci si concentra sulla parte di problem solving piuttosto che quella di problem posing. Pertanto si inizia già introducendo il modello che ha permesso di separare logicamente le variabili indipendenti dagli attributi nel Paragrafo 8.1. Il Paragrafo 8.2 presenta quindi l'algoritmo di induzione di PET incrementali.

8.1 Definizione del problema

Tentare di predire l'esito di una terapia, noto solo alla fine ma a varie fasi della stessa costringe a definire i particolari di un nuovo problema come spiegato nel capitolo precedente. La necessità scaturisce dal fatto che non si parla più di semplice predizione di un fenomeno immutabile, ma bensì di predizione di un processo in cui è possibile fare delle scelte che cambiano il risultato finale.

Il problema è definibile in forma generale ed è applicabile in qualsiasi contesto in cui si vorrebbe predire l'esito di un processo, quando esso è di tipo *categorico* o *nominale*, e lo si vorrebbe fare ad istanti diversi che precedono il momento in cui si sa realmente il risultato finale. La frequenza con cui si predice il suo esito non è dettata a priori, infatti è del tutto variabile il lasso di tempo tra una predizione e la successiva.

Per formalizzare il procedimento si ipotizzi di voler studiare un processo in atto sul sistema $\mathcal{S}$ in Figura 8.1: questa figura è la diretta generalizzazione della Figura 7.7. Il sistema possiede N ingressi $x_i(t)$, M stati $s_j(t)$ ed una sola uscita $y(t)$.

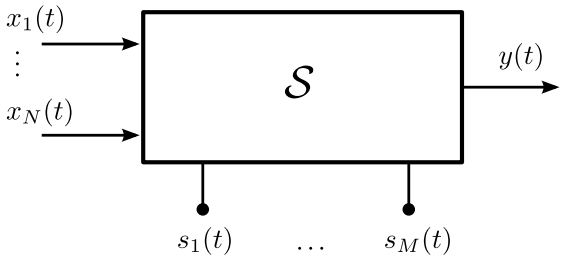


Figura 8.1: Sistema con N ingressi, M stati e una uscita

L'uscita del sistema $\mathcal{S}$ è la proprietà che si vuole studiare del processo. Essa è di tipo categorico e dipende dal tempo. Nel caso del problema in esame in questa tesi l'uscita è il fatto di essere LTR o non essere LTR. Naturalmente dipende dal tempo perché è nota con certezza solo alla fine del processo di guarigione, sia esso t_f. Gli ingressi $x_i(t)$ sono i veri attuatori del processo e possono essere fatti variare nel tempo a discrezione dell'utente. Gli stati $s_j(t)$ invece racchiudono in loro l'attività passata del sistema e sono soggetti anch'essi ad evolversi nel tempo. Questa è la diretta generalizzazione del caso medico in esame: gli ingressi, attuatori del processo di guarigione, sono i farmaci IFN e RBV; gli stati sono tutti i parametri clinici che possono valutare l'andamento della terapia.

Supponendo di avere effettuato n realizzazioni di un particolare processo, i dati a disposizione devono permettere di costruire più training set $T^{(t_k)}$ di taglia n ognuno relativo all'istante temporale t_k. Si ipotizzi inoltre di voler studiare l'esito del processo a q istanti temporali distinti, i training set per ogni istante temporale t_k sono così definiti:

$$\begin{aligned} T^{(t_0)} &\text{ definito sullo schema } \mathcal{A}^{(t_0)} \cup C \\ &\vdots \\ T^{(t_k)} &\text{ definito sullo schema } \mathcal{A}^{(t_k)} \cup C \\ &\vdots \\ T^{(t_{q-1})} &\text{ definito sullo schema } \mathcal{A}^{(t_{q-1})} \cup C \\ &\text{con } t_0 < t_1 < \cdots < t_{q-1} \end{aligned}$$

Gli attributi in $\mathcal{A}^{(t_k)}$ con $t_0 < t_1 < \cdots < t_{q-1}$ e $t_k \in \mathbb{R}^+$ sono attributi il cui valore è stato valutato studiando gli ingressi $x_i(t)$ e gli stati $s_j(t)$ del sistema $\mathcal{S}$ nella finestra temporale $[0, t_k]$. C è l'attributo target che rappresenta il valore di $y(t_f)$ dove $t_f \geq t_{q-1}$.

Molti degli attributi di $\mathcal{A}^{(t')}$ possono appartenere anche ad $\mathcal{A}^{(t'')}$ se $t' < t''$ ma *non* è vero in generale $\mathcal{A}^{(t')} \subseteq \mathcal{A}^{(t'')}$ in quanto gli attributi relativi a un test istantaneo vengono comunemente utilizzati per la predizione agli istanti successivi, invece quelli ottenuti combinando dati relativi all'intervallo temporale $[0, t']$ molto probabilmente devono essere ricalcolati nell'intervallo $[0, t'']$. Per rendere chiaro questo fatto si fornisce l'esempio 8.1.

Esempio 8.1. Si supponga di voler estrarre delle feature dalle n realizzazioni di un segnale $x(t)$. Ipotizzando di voler utilizzare questi segnali per predire l'esito di un processo all' istante t_1 è possibile procedere ad esempio in due modi:

1. Valutare il valore istantaneo $x^{(i)}(t_1)$ per $1 \leq i \leq n$, e produrre l'attributo A;
2. Valutare il *valore medio* assunto da $x^{(i)}(t)$ per $1 \leq i \leq n$ nell'intervallo $[0, t_1]$, e produrre l'attributo B.

I due attributi vengono aggiunti ad $\mathcal{A}^{(t_1)}$. L'attributo A può essere aggiunto ad $\mathcal{A}^{(t_2)}$. L'attributo B, per venire utilizzato coerentemente all'istante t_2, dovrebbe essere ricalcolato valutando il valore medio dei segnali $x^{(i)}(t)$ nell'intervallo $[0, t_2]$. Quindi $B \notin \mathcal{A}^{(t_2)}$, perciò $\mathcal{A}^{(t_1)} \nsubseteq \mathcal{A}^{(t_2)}$. ▪

Ad esempio nel caso medico esposto nel Capitolo 7, gli attributi relativi alle viremie di istanti temporali precedenti vengono aggiunti al training set corrente, ma il valore medio delle dosi assunte fino all'istante in questione deve essere ricalcolato.

I q training set T così definiti possono essere benissimo utilizzati come input del Problema di classificazione 3.1. Questo è stato il primo approccio risolutivo utilizzato per predire la risposta alla malattia. Nel particolare si sono ottenuti q alberi decisionali $M^{(t_k)}$ i cui attributi $\mathcal{A}^{(t_k)}$ erano i soli parametri clinici del paziente. L'insieme delle classi era invece $\Gamma = \{\text{“LTR”}, \text{“non LTR”}\}$.

Non vi è una separazione netta, in base agli attributi clinici, tra la classe “LTR” e “non LTR”. Per questo motivo si è deciso di costruire un PET per ogni training set $T^{(t_k)}$. La soluzione appena introdotta ha ancora come risultato la produzione di q modelli da utilizzare per la predizione dell'esito di un nuovo processo giunto solamente al tempo t_k e quindi non ancora completato.

Visto che il risultato del processo è direttamente indotto dagli *ingressi* del sistema è possibile studiarli separatamente dagli *stati* per vedere che modiche apportano all'esito finale. In particolare, giunti all'istante t_k, si ha ancora la possibilità di modificare gli ingressi del sistema $x_i(t)$ nella finestra (t_k, t_f). È doveroso puntualizzare che lo studio degli *stati* nel futuro, cioè relativamente alla finestra temporale (t_k, t_f), non è di alcun supporto alle decisioni in quanto comunque non si avrebbe la possibilità di andare a intervenire su di essi. Nel caso in esame in questa tesi, dopo un dato istante t_k si ha la possibilità di modificare la somministrazione di IFN e RBV (ingressi), ma solo di osservare i parametri clinici (stati) senza poterli modificare direttamente.

Si vuole allora introdurre un'altra definizione training set che tenga conto di questi aspetti:

Definizione 8.1. *Un training set con dati temporali $T^{(t_k)}$, relativo all'istante t_k, è un training set definito sullo schema $\mathcal{A}^{(t_k)} \cup C$. Gli attributi in $\mathcal{A}^{(t_k)}$ sono partizionabili in due insiemi:*

(i) $\mathcal{A}_p^{(t_k)}$ (attributi passati e fissi): insieme di attributi il cui valore è stato valutato studiando gli ingressi $x_i(t)$ e gli stati $s_j(t)$ di un sistema $\mathcal{S}$ nella finestra temporale $[0, t_k]$.

(ii) $\mathcal{A}_f^{(t_k)}$ (attributi futuri e modificabili): insieme di attributi il cui valore è stato valutato studiando i soli ingressi $x_i(t)$ di un sistema $\mathcal{S}$ nella finestra temporale $[0, t_f)$.

C è l'attributo target che rappresenta il valore di $y(t_f)$.

La definizione è quanto più possibile generale, infatti si specifica che tra gli attributi *futuri e modificabili* vi è la possibilità di includere attributi il cui valore è calcolato in base a tutta la durata del processo invece che solo quelli relativi alla finestra temporale (t_k, t_f). Il solo fatto di produrre una feature studiando un segnale per istanti successivi all'istante t_k dato classifica quell'attributo come futuro. Questa è la generalizzazione dello studio della dose media, e quindi in rapporto con la dose standard, del Capitolo 7.

Ipotizzando di voler predire l'esito di un nuovo processo a q istanti di tempo distinti vengono prodotti q training set temporali:

$$\begin{aligned} T^{(t_0)} &\text{ definito sullo schema } \mathcal{A}_p^{(t_0)} \cup \mathcal{A}_f^{(t_0)} \cup C \\ &\vdots \\ T^{(t_k)} &\text{ definito sullo schema } \mathcal{A}_p^{(t_k)} \cup \mathcal{A}_f^{(t_k)} \cup C \\ &\vdots \\ T^{(t_{q-1})} &\text{ definito sullo schema } \mathcal{A}_p^{(t_q)} \cup \mathcal{A}_f^{(t_q)} \cup C \\ &\text{con } t_0 < t_1 < \cdots < t_{q-1} \end{aligned}$$

È necessario non confondere il significato intrinseco degli attributi *passati e fissi* e di quelli *futuri e modificabili*, utilizzarli indistintamente per costruire un modello predittivo potrebbe portare dei problemi. Si consideri il seguente esempio pratico, relativo alla scelta della dose ottimale per curare un paziente, per intuire quali siano i problemi. Si vuole costruire un PET utilizzando indistintamente gli attributi passati da quelli futuri. È naturalmente possibile che un attributo relativo al futuro venga scelto prima di un attributo passato nell'albero. Si osservi l'esempio in Figura 8.2 dove è rappresentato un PET per la stima della probabilità di ottenere la LTR o meno. Tale PET utilizza come attributo

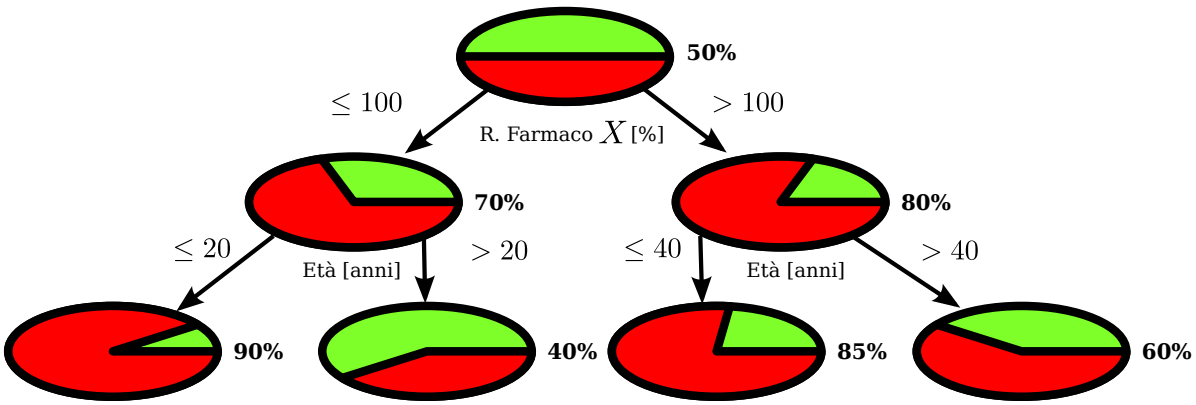


Figura 8.2: PET che usa indistintamente attributi passati e futuri

futuro "R. Farmaco X": il rapporto tra la dose media che il paziente andrà a fare e la

dose standard della terapia. L'attributo passato utilizzato è "Età". Il medico potrebbe utilizzare questo modello per scegliere la dose ottimale con cui trattare un nuovo paziente y con età inferiore o uguale ai 20 anni (y["Età"] ≤ 20). Egli dovrebbe valutare il valore degli attributi del paziente in esame e percorrere l'albero dalla radice coerentemente con i risultati dei test sugli archi. Il valore dell'attributo futuro "R. Farmaco X" per il paziente y non però noto, gli si porrebbe di fronte la scelta di decidere egli stesso il percorso da seguire nel PET e di conseguenza il valore di tale rapporto per il futuro. Una scelta *localmente* ottima di tale rapporto, fatta osservando il consiglio del nodo radice, lo potrebbe portare ad ottenere una probabilità di risposta minore di quella massima ottenibile per il paziente y. Tale nodo indica che il rapporto della dose dovrebbe essere superiore all'100% per avere più probabilità di essere LTR , pari al 80%, rispetto a un rapporto $\leq 100\%$ che stima una probabilità del 70%. La prima scelta però porterebbe il paziente y ad avere una probabilità di essere LTR dell'85% visto l'età inferiore o uguale ai 20 anni, la seconda invece del 90%. È allora più conveniente scegliere una dose il cui rapporto è ≤ 100 per garantire al paziente y una probabilità maggiore di essere LTR.

L'esempio fa intuire come possa essere operativamente complicato scegliere la dose ottimale per un nuovo paziente se vengono mescolati nodi associati a un test su attributo futuro e quelli associati a test su attributi passati. Il medico dovrebbe percorrere tutto il sottoalbero relativo alle dose scelta e confrontarlo con quello della dose alternativa per conoscere il dosaggio realmente migliore. Costruendo un albero sviluppato inizialmente con nodi relativi ad attributi passati e solo poi con nodi associati ad attributi futuri permette di stimare più efficacemente la probabilità di risposta del nuovo paziente. Una volta identificato usando gli attributi passati il nodo in cui finisce il soggetto, sia esso g, è possibile osservare il sottoalbero sviluppato con i parametri futuri che ha per radice tale nodo e identificare la foglia con maggiore probabilità, sia essa u. Le scelte future ottime per il PET ottenuto saranno quindi indicate dal cammino da g ad u.

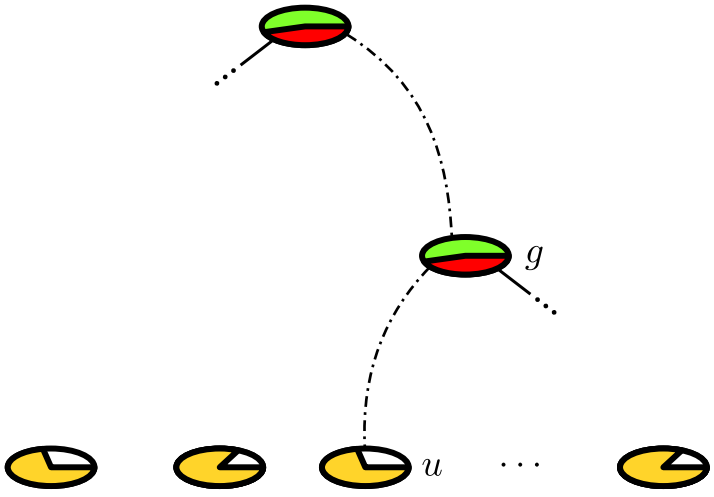


Figura 8.3: Scelte future ottime

Questo è il motivo per cui si costruiscono alberi *incrementali* sviluppati inizialmente sui

parametri passati e poi su quelli futuri. La definizione formale e l'algoritmo di induzione sono presentati nel prossimo paragrafo.

8.2 PET incrementale

Come discusso nel paragrafo precedente, un PET incrementale è utile per la previsione dell'esito di un processo a un dato istante t_k ed è applicabile a un processo il cui risultato è categorico. Un PET incrementale permette di stimare la probabilità delle possibili uscite del processo basandosi prima sull'osservazione degli attributi passati e poi su quelli futuri per fornire un supporto alle scelte che si possono fare al tempo t_k.

La definizione formale di un PET incrementale viene di seguito introdotta:

Definizione 8.2. *Un PET incrementale per un training con dati temporali $T^{(t_k)}$ è un PET costruito in due fasi:*

(i) Viene costruito un PET su $T^{(t_k)}$ considerandolo sullo schema $\mathcal{A}_p^{(t_k)} \cup C$;

(ii) Per ogni nodo foglia u del PET ottenuto, a cui è associato il sottoinsieme di record $T_u^{(t_k)}$, viene costruito un PET, con radice u, su $T_u^{(t_k)}$ considerandolo sullo schema $\mathcal{A}_f^{(t_k)} \cup C$.

La Denizione 8.2 evidenzia come un PET incrementale sia di fatto costruito in due fasi. La prima che lo sviluppa rispetto gli attributi passati e la seconda rispetto a quelli futuri. L'algoritmo deve essere parametrizzabile per poter scegliere la profondità con cui l'albero si sviluppa rispetto al passato e quanto si sviluppa rispetto al futuro. Questo viene fatto implicitamente settando la cardinalità minima dei nodi foglia sviluppati con gli attributi passati e quella minima dei nodi foglia sviluppati per il futuro. Il risultato si ottiene modificando l'algoritmo di induzione di alberi decisionali di Hunt presentato nell'Algoritmo 5.1. Lo pseudocodice è presentato nell'Algoritmo 8.1. La prima chiamata dell'algoritmo viene fatta con il training set temporale $T^{(t_k)}$ relativo all'istante t_k.

La funzione StoppingConditionPast permette di definire quando fermare la crescita dell'albero rispetto agli attributi passati $\mathcal{A}_p^{(t_k)}$. Come per l'algoritmo classico la scelta è fatta in base alla cardinalità dell'insieme di record $E^{(t_k)}$ associato ai nodi. Il valore minimo di taglia si definisce N_p. L'altra condizione è che non si riescano a trovare attributi di $\mathcal{A}_p^{(t_k)}$ che permettano di ottenere split che aumentano la purezza.

Una volta finito il processo di crescita rispetto agli attributi passati viene chiamata la funzione TreeGrowthFuture per sviluppare ulteriormente l'albero rispetto agli attributi futuri $\mathcal{A}_f^{(t_k)}$. La condizione di stop in questo caso è definita da StoppingConditionFuture che, oltre a fermare l'induzione nel caso non si riescano a trovare split interessanti sugli attributi $\mathcal{A}_f^{(t_k)}$, determina lo stop nel caso la cardinalità dei record dei nodi foglia sia inferiore a N_f. Alle foglie è stimata la distribuzione di probabilità delle classi con la funzione EstimateClassProbabily utilizzata nel Capitolo 6 per i PET classici.

Si noti che se $N_p/N_f < 2$ non verranno mai costruiti sottoalberi relativi al futuro per nodi terminali u dell'albero indotto con i soli attributi passati se $|E_u^{(t_k)}| = N_p$.

A riguardo della valutazione della qualità di un PET incrementale si vuole dire che la stima di probabilità offerta dai nodi foglia può essere valutata con la tecnica della ROC curve, come per i PET classici.

Algoritmo 8.1 Algoritmo per l'induzione di un PET incrementale

IncrementalTreeGrowth($E^{(t_k)}, \mathcal{A}_p^{(t_k)}, \mathcal{A}_f^{(t_k)}, C$)

if StoppingConditionPast($E^{(t_k)}, \mathcal{A}_p^{(t_k)}, C$) **then**
    Crea una foglia *leaf* associata ad $E^{(t_k)}$
    leaf.label $\leftarrow$ TreeGrowthFuture($E^{(t_k)}, \mathcal{A}_f^{(t_k)}, C$)
else
    Crea un nodo *root*
    root.test_cond $\leftarrow$ FindBestSplit($E^{(t_k)}, \mathcal{A}_p^{(t_k)}, C$)
    $V \leftarrow \{v : v$ possibile risultato del test *root.test_cond*$\}$
    for each $v \in V$ **do**
        $E_v^{(t_k)} \leftarrow \{e \in E^{(t_k)} :$ *root.test_cond*$(e) = v\}$
        child $\leftarrow$ IncrementalTreeGrowth($E_v^{(t_k)}, \mathcal{A}_p^{(t_k)}, \mathcal{A}_f^{(t_k)}, C$)
        Aggiungi *child* come foglia di *root* con un arco etichettato v
return *root*

TreeGrowthFuture($E^{(t_k)}, \mathcal{A}_f^{(t_k)}, C$)

if StoppingConditionFuture($E^{(t_k)}, \mathcal{A}_f^{(t_k)}, C$) **then**
    Crea una foglia *leaf* associata ad $E^{(t_k)}$
    leaf.label $\leftarrow$ EstimateClassProbabily($E^{(t_k)}, \mathcal{A}_f^{(t_k)}, C$)
else
    Crea un nodo *root*
    root.test_cond $\leftarrow$ FindBestSplit($E^{(t_k)}, \mathcal{A}_f^{(t_k)}, C$)
    $V \leftarrow \{v : v$ possibile risultato del test *root.test_cond*$\}$
    for each $v \in V$ **do**
        $E_v^{(t_k)} \leftarrow \{e \in E^{(t_k)} :$ *root.test_cond*$(e) = v\}$
        child $\leftarrow$ TreeGrowthFuture($E_v^{(t_k)}, \mathcal{A}_f^{(t_k)}, C$)
        Aggiungi *child* come foglia di *root* con un arco etichettato v
return *root*

Capitolo 9

Risultati sperimentali

In questo capitolo vengono presentati i risultati sperimentali ottenuti. Nel Paragrafo 9.1 vengono introdotti i tipi di test effettuati e confrontati i risultati che si sono ottenuti parametrizzando diversamente l'algoritmo di costruzione dei PET incrementali. Nel Paragrafo 9.2 vengono proposti e discussi alcuni PET incrementali costruiti per stimare la probabilità di essere LTR o meno di un nuovo paziente prima del trattamento, al primo mese e al terzo mese di terapia.

9.1 Test effettuati

Come discusso nel Capitolo 7, i pazienti scelti sono infetti da HCV genotipo 1 e 4 che sono stati trattati con la terapia standard di Peg-IFN in associazione a RBV. I modelli ottenuti allo stato basale del paziente, al primo mese e al terzo di terapia permettono di stimare la probabilità di ottenere la risposta sostenuta per un nuovo paziente osservando la storia clinica trascorsa fino ad un certo istante. Inoltre essi permettono di vedere come possa variare tale probabilità rispetto a un possibile cambiamento di dose fatto in quel momento.

In questo studio si sono utilizzati come parametri futuri il rapporto della dose di farmaco rispetto a quella standard per IFN e RBV. Questo ha permesso di utilizzare i pazienti che hanno assunto Peg-IFN alfa-2a e Peg-IFN alfa-2b in modo indistinto. Quelli passati sono stati i parametri clinici del paziente e le dosi assunte fino all'istante in cui si va ad osservare il modello. Il risultato desiderato si è ottenuto modificando *J48* di Weka [34], che è l'implementazione in Java del famoso C4.5 più volte citato in questa tesi. L'implementazione segue la linea di principio dell'Algoritmo 8.1 per costruire un PET incrementale. In fase di esecuzione viene permessa la scelta delle seguenti opzioni:

- `-M` <N_p>
- `-MF` <N_f>
- `-A`

L'ultima opzione permette di scegliere se la stima di probabilità verrà fatta con il metodo della massima verosimiglianza o con la correzione di Laplace. Gli alberi incrementali per stima di probabilità ottenuti vengono sviluppati inizialmente con i parametri passati e solo successivamente con i parametri futuri. Il passaggio dalla fase iniziale a quella successiva è determinato in primo luogo dal valore di N_p, che indica la cardinalità minima a cui possono arrivare i nodi sviluppati nella prima fase. L'albero si sviluppa rispetto agli attributi passati finché raggruppa i pazienti in sottoinsiemi non più piccoli di N_p. Questi sottoinsiemi vengono poi suddivisi rispetto ai parametri futuri. La profondità dell'albero relativamente ai parametri futuri è determinata dal numero minimo (N_f) di pazienti che si ammettono ad un nodo foglia. È anche possibile che un sottoinsieme di pazienti non riesca ad essere suddiviso con un test relativo a un attributo passato anche se di cardinalità maggiore di N_p. A questo punto vengono utilizzati gli attributi futuri per suddividere ulteriormente i pazienti. Con diversi valori di N_p e N_f si ottengono risultati differenti. Scegliere valori piccoli per tali parametri induce lo sviluppo più in profondità dell'albero. Si ricorda che tutte le suddivisioni mirano ad isolare con un test i pazienti non LTR dagli LTR. Il valore di N_p viene scelto con l'opzione `-M` e il valore di N_f con l'opzione `-MF`.

La stima di probabilità dei modelli ottenuti è stata valutata con le ROC curve. Le ROC curve valutano infatti la capacità del modello di discernere tra pazienti LTR e pazienti non LTR con le probabilità. Quelle presentate in questo capitolo sono state ottenute mediando 10 curve ottenute con una 10-fold cross-validation. Il procedimento per farlo è spiegato nel Paragrafo 6.3.1.

La selezione dei parametri N_p e N_f permette di ottenere modelli radicalmente diversi: con valori piccoli di tali parametri gli alberi sono molto più corposi a differenza di quelli che si otterrebbero con valori grandi. Vengono quindi presentati i test effettuati variando i valori di N_p e N_f in Tabella 9.1 e senza settare l'opzione `-A`, quindi ottenendo una stima di probabilità alle foglie di massima verosimiglianza. In questa tabella si nota come i modelli migliorino al passare del tempo. I modelli costruiti al primo mese di terapia sono migliori di quelli relativi allo stato basale del paziente così come quelli del terzo mese sono migliori del primo. Questo fatto è dovuto alla maggiore quantità di informazioni con cui vengono costruiti i modelli a istanti temporali avanzati di terapia.

Osservando un singolo istante temporale si nota come parametri N_p e N_f a valori più piccoli permettano di ottenere AUC più elevate. Questo risultato è coerente con quanto detto per la teoria dei PET al Capitolo 6: più un albero è sviluppato e migliore è la stima di probabilità ottenuta. Naturalmente, a causa di fluttuazioni statistiche valori troppo bassi dei parametri portano ad ottenere alberi poco realistici.

Osservando le ROC curve in Figura 9.1 è possibile notare il miglioramento che si ottiene con valori bassi di N_p e N_f. Le ROC disegnate sono tutte relative allo stato basale del paziente. Il miglioramento è ancora più visibile se si variano i parametri per costruire un modello al terzo mese. Il confronto può essere fatto osservando la Figura 9.2. Il miglioramento più marcato nella variazione dei parametri al terzo mese è dovuto alla maggiore disponibilità di informazioni cliniche sul paziente. Le stime di probabilità ne traggono giovamento.

Tempo	N_p	N_f	Rapporto	AUC
Basale	70	40	1.75	0.608
1° mese	70	40	1.75	0.771
3° mese	70	40	1.75	0.819
Basale	60	30	2.00	0.593
1° mese	60	30	2.00	0.765
3° mese	60	30	2.00	0.857
Basale	50	20	2.50	0.635
1° mese	50	20	2.50	0.765
3° mese	50	20	2.50	0.860
Basale	40	20	2.00	0.644
1° mese	40	20	2.00	0.753
3° mese	40	20	2.00	0.885
Basale	40	15	2.67	0.640
1° mese	40	15	2.67	0.755
3° mese	40	15	2.67	0.885
Basale	30	15	2.00	0.668
1° mese	30	15	2.00	0.758
3° mese	30	15	2.00	0.895
Basale	20	10	2.00	0.650
1° mese	20	10	2.00	0.779
3° mese	20	10	2.00	0.889

Tabella 9.1: Test dei parametri

Un'analisi qualitativa della ROC curve rivela che al basale la scelta accorta dei parametri è in grado di migliorare sensibilmente le stime di probabilità per gli LTR. Si noti infatti la differenza tra la curva con parametri 70 e 40 e quello con 30 e 15: l'ultima si porta leggermente verso la retta di equazione TPR = 1. Al terzo mese l'accuratezza nella predizione di un LTR è molto buona. La variazione nei parametri permette invece di ottenere migliori stime per i non LTR. Infatti le curve con N_p e N_f piccoli si schiacciano verso l'asse FPR = 0. Il rapporto N_p/N_f serve per valutare quanto un albero è sviluppato nel passato rispetto al futuro. Valori alti di rapporto indicano un albero molto sviluppato nei parametri futuri, valori bassi invece indicano il contrario. In questi test tale rapporto è sempre stato posto ad un valore minore di 3. L'informazione rispetto alle scelte future deve essere contenuta e per questo i rapporti sono bassi. Un albero eccessivamente sviluppato nei parametri futuri, essendo essi solo due, può portare solo a split spuri. Vengono definiti split spuri le suddivisioni indotte dai nodi interni del PET ottime per il gruppo di pazienti in esame ma di scarsa qualità per un soggetto con HCV qualsiasi.

Per mostrare l'effettiva validità di questo approccio per stimare la probabilità di risposta sostenuta alla terapia per un nuovo paziente vengono proposti tre alberi: uno allo stato basale, uno al primo mese di terapia e uno al terzo. Questi sono stati selezionati tra quelli di test in base alla misura AUC e a quanto fossero conformi alle aspettative mediche.

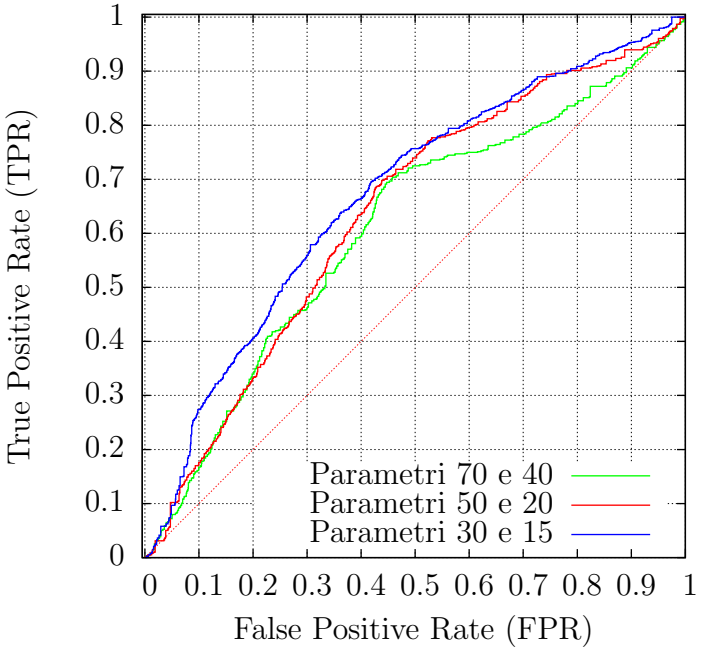


Figura 9.1: ROC curve disegnate per un modelli costruiti per lo stato basale del paziente con diversi parametri N_p e N_f

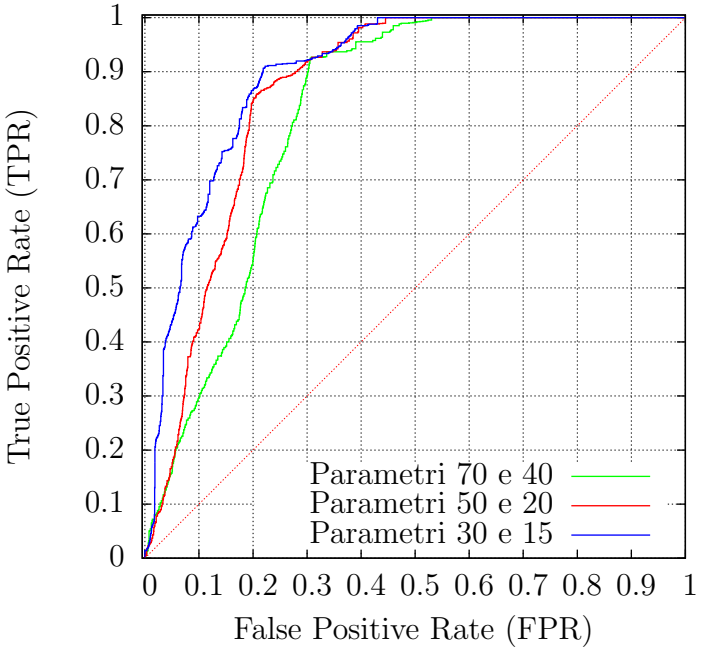


Figura 9.2: ROC curve disegnate per un modelli costruiti per il terzo mese di terapia con diversi parametri N_p e N_f

9.2 Modelli scelti dai medici

In questo paragrafo vengono mostrati tre alberi selezionati dai medici epatologi esperti nel settore tra quelli ottenuti con i test. Le buone misure di AUC, soprattutto al primo e terzo mese di terapia, permettono di farne un reale utilizzo in pratica.

Si ricordi che stime di probabilità fatte nei nodi con elevata cardinalità sono da considerarsi più attendibili di quelle fatte su gruppi piccoli di pazienti. Se fosse stato utilizzato l'approccio Bayesiano alla stima, la differenza si sarebbe rivelata in termini di intervallo di confidenza: più ampio per le stime fatte su pochi pazienti e più stretto per quelle fatte su molti. I modelli più compatti presentano nodi foglia a cui sono associati insiemi con elevato numero di pazienti, a differenza dei PET incrementali meno compatti le cui foglie sono associate ad insiemi a bassa cardinalità. Sebbene le stime di probabilità siano fatte con minor confidenza, sono proprio gli ultimi i modelli da cui si possono estrarre molte più informazioni. Vengono infatti mostrati legami non banali tra le variabili che difficilmente si sarebbero potuti ipotizzare.

Nei sottoparagrafi successivi vengono confrontati inizialmente due modelli allo stato basale per far notare la differenza tra un albero più compatto e uno meno compatto. Di seguito si discute la selezione di altri due modelli, uno relativo al primo mese di terapia e uno relativo al terzo. Gli ultimi sono stati costruiti con N_p e N_f bassi, il che ha permesso di ottenere alberi molto profondi.

9.2.1 Modello allo stato basale

In Tabella 9.2 vengono elencati i parametri utilizzati per la costruzione del modello relativo allo stato basale del paziente. Viene proposto un albero molto compatto costruito con N_p e N_f rispettivamente a 70 e a 40 e un secondo più sviluppato con i parametri 50 e 20. I due alberi sono visibili in Figura 9.3 e 9.4. Questi ottengono rispettivamente un valore di AUC di 0.608 e 0.635 ottenuti calcolando l'area sottesa dalle ROC curve in Figura 9.1.

Passati e Fissi
Sesso [Maschio/Femmina], Genotipo [HCV1/HCV4], Categoria [Epatite Cronica C/Cirrosi C], Durata malattia [$mesi$], Peso [kg], IMC [kg/m^2], Età [$anni$], Viremia basale [UI/mL], Emoglobina basale [g/L], Globuli Bianchi basale [$n. \cdot 10^9/L$], Piastrine basale [$n. \cdot 10^9/L$], AST basale [U/L], ALT basale [U/L], GGT basale [U/L]
Futuri e Manipolabili
R. IFN [%], R. RBV [%]

Tabella 9.2: Parametri utilizzati per il modello costruito allo stato basale del paziente

In entrambi i casi i modelli catturano i parametri predittivi che vengono considerati in letteratura medica come i più importanti. La numerosità più ridotta dei sottoinsiemi di pazienti che può ottenere il secondo gli permette di identificare la *viremia basale* come primo fattore e gli permette inoltre di presentare molte più informazioni rispetto al primo, pertanto è da considerarsi migliore: la scelta del modello per lo stato basale deve quindi indirizzarsi su di esso. Il cut-off trovato sulla viremia basale si aggira sui 5 logaritmi in

base 10 che è proprio la soglia che i medici epatologi si sarebbero aspettati. Un altro fattore legato con la difficoltà di ottenere la risposta sostenuta è stato l'essere cirrotici o meno. Le probabilità stimate di riuscita del trattamento per tali pazienti sono sempre molto ridotte. Una volta discriminato in base ai fattori maggiormente legati alla risposta sostenuta vengono scelti i parametri *età* e *peso*. Il risultato è coerente: età avanzate e pesi elevati sono avversi alla possibilità di ottenere la LTR.

Quando possibile il modello produce un consiglio sul dosaggio di IFN e RBV da seguire in terapia. Si osservi a questo proposito la Figura 9.3. I pazienti con epatite cronica la cui viremia è molto elevata (sopra le 830000 UI/mL) è bene che assumano il 99% della dose standard di IFN. Le probabilità stimate di rispondere alla terapia sono del 28% per chi fosse costretto a sottodosaggi di IFN e del 52% per chi riuscisse a seguire il dosaggio indicato.

Le informazioni prodotte dai PET incrementali possono essere mal interpretate se non osservate in dettaglio e con il supporto di esperti del settore. A causa di fluttuazioni statistiche è possibile che in certi casi l'albero produca split spuri. Ad esempio, si prenda in considerazione lo split sulla dose di RBV consigliata in Figura 9.4. L'albero sembra proporre che una dose di RBV inferiore al 99% di quella standard sortisca risultati migliori in terapia di una dose superiore al 99%. Le stime di probabilità di ottenere la LTR sono il 74% per chi assume meno dell 99% di quella standard e il 49% per quelli che assumono più del 99%. Si ricorda che le probabilità indicate nei nodi sono stime fatte sul campione di pazienti da cui è stato indotto l'albero. Stime fatte su campioni più piccoli sono meno precise di quelli fatte su grandi campioni, e l'albero di Figura 9.4 suddivide i pazienti in gruppi più piccoli di quello di Figura 9.3. Tale nodo spurio è imputabile ad una fluttuazione statistica, comunque sia apporta informazione al medico epatologo esperto del settore. Le linee guida terapeutiche permettono infatti di scalare la dose di RBV per pazienti giovani con basso peso corporeo. Tale nodo nel modello è infatti estrapolato da pazienti con peso $\leq 74\ kg$ e età $\leq 51\ anni$ come si può notare sempre in Figura 9.4. Solo con l'informazione aggiuntiva dell'esperto del settore è possibile valutare quel nodo oggettivamente: i pazienti leggeri non subiscono un cambiamento radicale nella probabilità di riuscita della terapia sebbene la dose di RBV gli venga scalata. È quindi solo con occhio critico che ci si può approcciare correttamente al modello e che si può usufruire della maggior quantità di informazione fornita da alberi profondi.

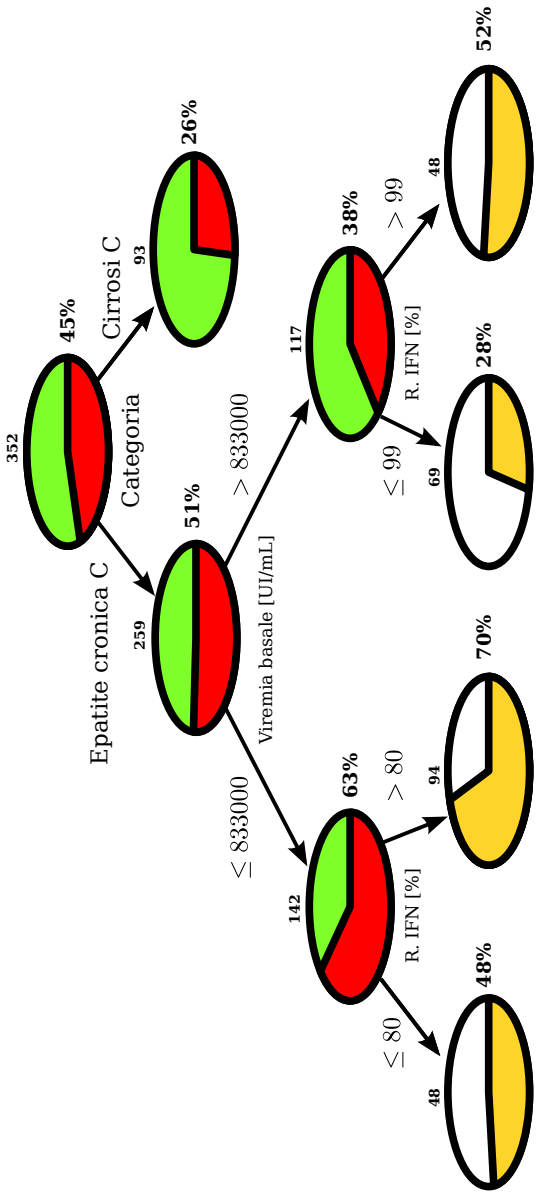


Figura 9.3: Modello compatto relativo allo stato basale del paziente

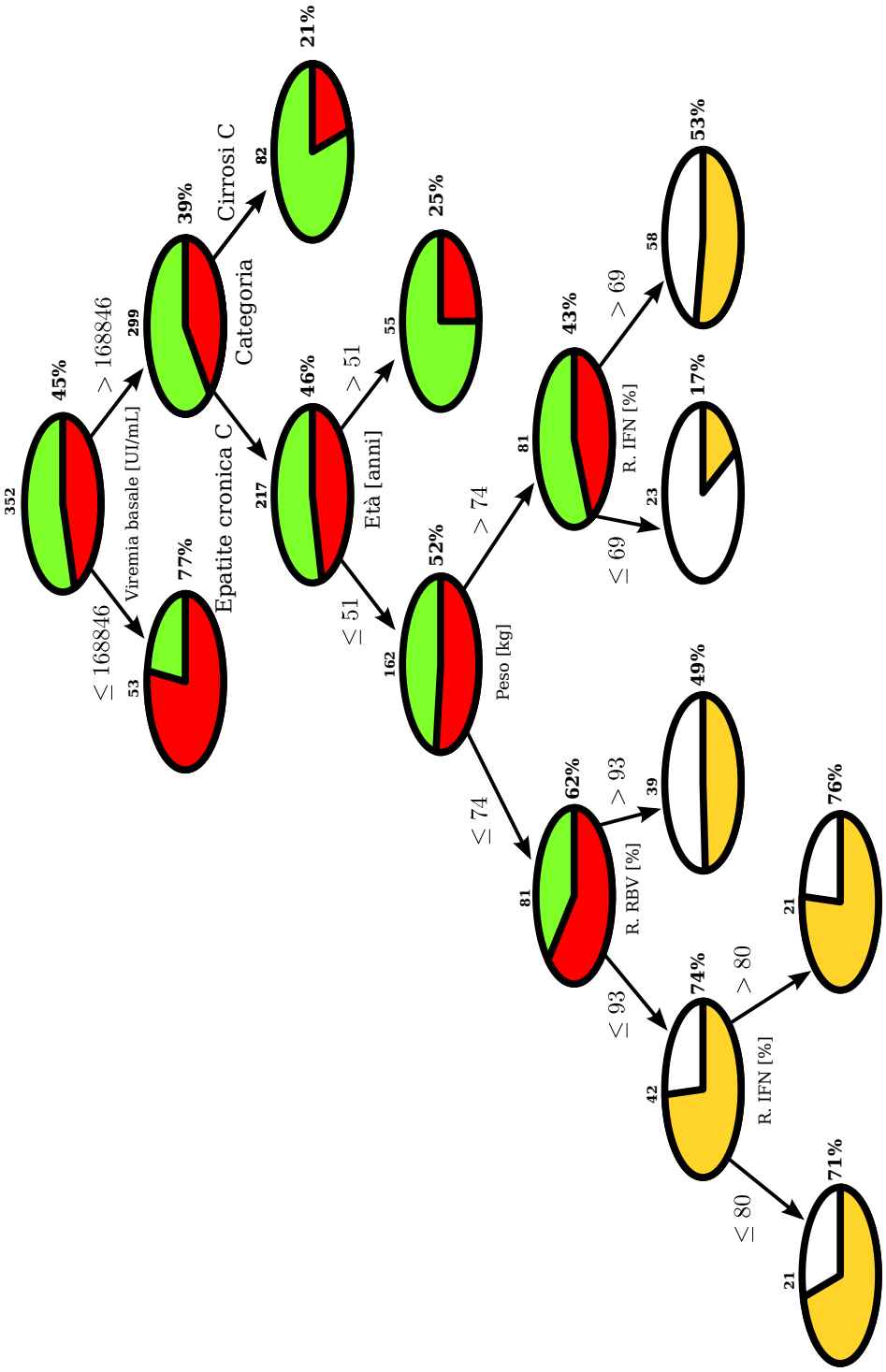


Figura 9.4: Modello più sviluppato relativo allo stato basale del paziente

9.2.2 Modello al primo mese di terapia

In questo paragrafo viene presentato un modello ottenuto con i parametri $N_p = 20$ e $N_f = 10$ relativo al primo mese di terapia. L'albero è molto sviluppato ed è presentato in Figura 9.5. La ROC curve relativa è disegnata in Figura 9.6, ad essa è associato un valore di AUC di 0.779. I parametri utilizzati per costruire l'albero sono elencati per completezza in Tabella 9.3. Ancora una volta il modello utilizza i parametri più predittivi segnalati

Passati e Fissi
Sesso [Maschio/Femmina], Genotipo [HCV1/HCV4], Categoria [Epatite Cronica C/Cirrosi C], Durata malattia [$mesi$], Peso [kg], IMC [kg/m^2], Età [$anni$], Viremia basale [UI/mL], Viremia 1^a settimana [UI/mL], Viremia 1° mese [UI/mL], Emoglobina basale [g/L], Globuli Bianchi basale [$n. \cdot 10^9/L$], Piastrine basale [$n. \cdot 10^9/L$], AST basale [U/L], ALT basale [U/L], GGT basale [U/L], Emoglobina 1° mese [g/L], Globuli Bianchi 1° mese [$n. \cdot 10^9/L$], Piastrine 1° mese [$n. \cdot 10^9/L$], AST 1° mese [U/L], ALT 1° mese [U/L], GGT 1° mese [U/L], R. IFN prima 1°m. [%], R. RBV prima 1°m. [%]
Futuri e Manipolabili
R. IFN dopo 1°m. [%], R. RBV dopo 1°m. [%]

Tabella 9.3: Parametri utilizzati per il modello costruito allo al primo mese di terapia

dalle linee guida. Non è banale la scelta della viremia al primo mese come nodo radice dell'albero: tale parametro viene scelto tramite ricerca esaustiva tra tutti i parametri a disposizione e selezionato come il migliore. Di secondaria importanza è il fatto di essere cirrotici o meno. Sia per i pazienti con carica virale bassa al primo mese ($\leq 9456\ UI/mL$) che per quelli con viremia più alta ($> 9456\ UI/mL$) la probabilità di ottenere la LTR è ridotta nei casi di cirrosi. L'albero sviluppato con i parametri 20 e 10 è più articolato e può apportare una grossa quantità di informazioni al medico. In ultima analisi può apportare anche informazioni sulla cinetica proponendo split sulla viremia basale. Una volta osservata la carica virale al primo mese, il PET in Figura 9.5 può infatti discriminare i pazienti in base a quello che era la carica virale prima del trattamento.

In conclusione si noti come sia vasta la proposta sulle dosi di IFN e RBV da assumere dopo il primo mese. La terapia è infatti in una fase non troppo avanzata in cui è ancora possibile apportarvi cambiamenti.

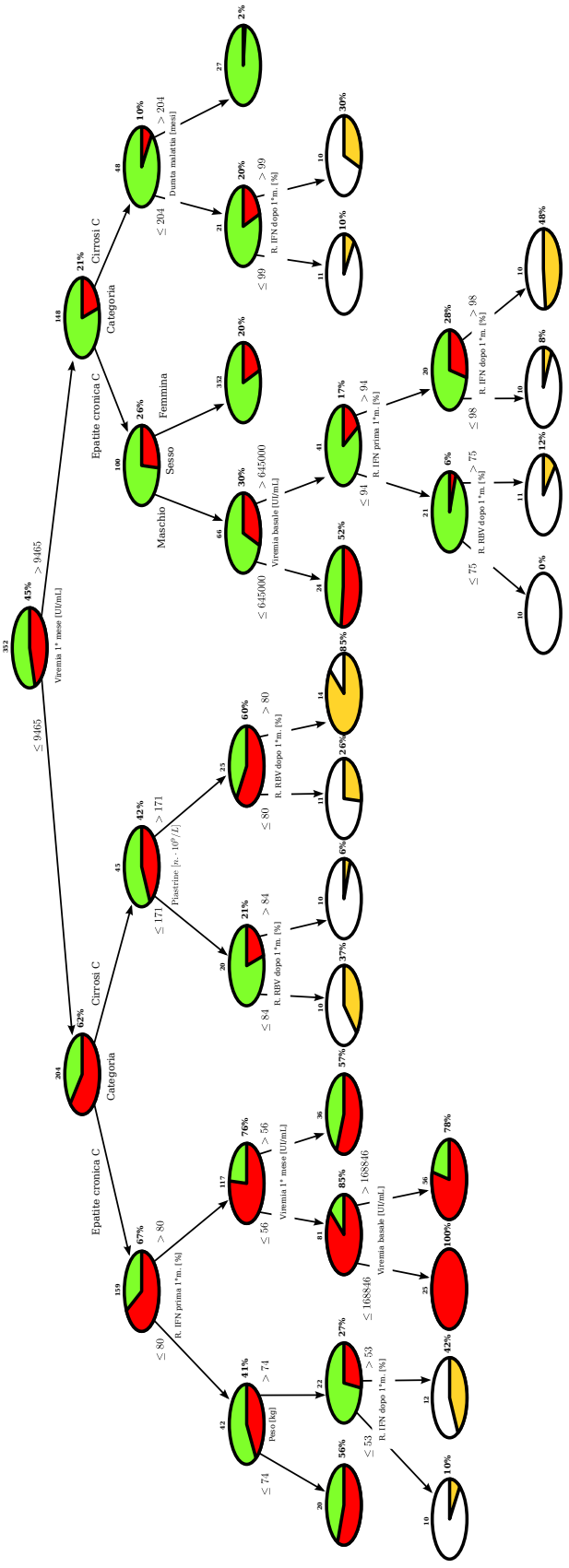


Figura 9.5: Modello relativo al primo mese di terapia

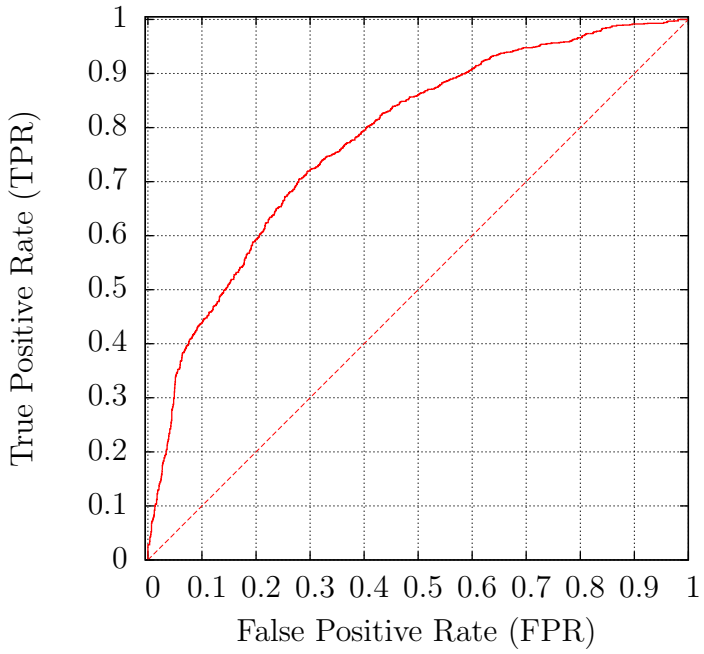


Figura 9.6: ROC curve disegnata per un modello costruito per il primo mese di terapia con N_p e N_f rispettivamente 20 e 10

9.2.3 Modello al terzo mese di terapia

In questo caso viene costruito un albero con i parametri N_p e N_f settati a 20 e a 10. L'AUC della ROC presentata in Figura 9.8 è ben 0.889. I parametri passati con cui è stato costruito sono relativi al trattamento fino al terzo mese, e sono mostrati in Tabella 9.4.

Passati e Fissi
Sesso [Maschio/Femmina], Genotipo [HCV1/HCV4], Categoria [Epatite Cronica C/Cirrosi C], Durata malattia [$mesi$], Peso [kg], IMC [kg/m^2], Età [$anni$], Viremia basale [UI/mL], Viremia 1^a settimana [UI/mL], Viremia 1° mese [UI/mL], Viremia 3° mese [UI/mL], Emoglobina basale [g/L], Globuli Bianchi basale [$n. \cdot 10^9/L$], Piastrine basale [$n. \cdot 10^9/L$], AST basale [U/L], ALT basale [U/L], GGT basale [U/L], Emoglobina 1° mese [g/L], Globuli Bianchi 1° mese [$n. \cdot 10^9/L$], Piastrine 1° mese [$n. \cdot 10^9/L$], AST 1° mese [U/L], ALT 1° mese [U/L], GGT 1° mese [U/L], Emoglobina 3° mese [g/L], Globuli Bianchi 3° mese [$n. \cdot 10^9/L$], Piastrine 3° mese [$n. \cdot 10^9/L$], AST 3° mese [U/L], ALT 3° mese [U/L], GGT 3° mese [U/L], R. IFN prima 3°m. [%], R. RBV prima 3°m. [%]
Futuri e Manipolabili
R. IFN dopo 3°m. [%], R. RBV dopo 3°m. [%]

Tabella 9.4: Parametri utilizzati per il modello costruito al terzo mese di terapia

Il terzo mese è, come più volte detto in questa tesi, un momento cruciale per capire quali sono le sorti del paziente. A seconda di quale sia la sua capacità di eradicazione virale, la probabilità di essere LTR o meno varia significativamente. Pertanto, l'albero di Figura 9.7 si spinge meno nel consiglio sulla dose futura (terzo mese in poi) di IFN o RBV,

rispetto alle indicazioni del PET al primo mese, ma decreta che sia prioritario valutare i parametri passati della terapia per stimare la probabilità di guarigione. Infatti, viene proposto un solo split su un attributo futuro e modificabile.

Per la discussione in oggetto si preferisce parlare dettagliatamente del sottoalbero destro della radice, che ha suscitato particolare interesse nei medici che lo hanno studiato. Tale sottoalbero è relativo ai pazienti che hanno viremia al terzo mese superiore le 10 UI/ml, quindi non hanno ottenuto la cEVR. Le linee guida internazionali prevedono di concludere la terapia di chi non ottiene la EVR, versione più debole di risposta rispetto alla cEVR. Le stesse linee guida indicano comunque che il non ottenere la cEVR è sintomo della poca efficacia del trattamento. L'albero mostra che l'incidenza di questi casi è molto elevata: sono 167 nel campione in esame. Se lo stop venisse esteso a tali pazienti si potrebbe negare la risposta sostenuta a ben l'8% di essi. Ma è possibile isolare quell'8% di soggetti? L'albero propone un metodo per farlo. I medici epatologi che hanno studiato il PET di Figura 9.7 hanno trovato quindi interessante che oltre la soglia delle 1890 UI/mL nessuno dei pazienti riesca ad ottenere la LTR. Il modo in cui vengono discriminati i pazienti con HCV-RNA positivo e test viremico quantitativo minore o uguale alle 1890 UI/mL deve essere motivo di studio ulteriore ed è molto interessante. La RBV è un forte emolitico e il medico epatologo è costretto a scalare la dose somministrata se il paziente anemizza. Tale effetto collaterale sembra essere un predittore della non LTR osservando il modello, ma è imputabile ad uno scalo di RBV avvenuto prima del terzo mese. Sotto questa ipotesi prende consistenza la scelta del modello di utilizzare il rapporto di RBV come attributo futuro per i pazienti anemici. Infatti un ulteriore scalo della dose di RBV successivo al terzo mese di trattamento potrebbe portare alla non guarigione del soggetto.

Ancora una volta il supporto del medico epatologo risulta necessario per la discriminazione di un nonsenso da una situazione clinica particolare. Il modello è infatti stato adattato alle esigenze dei medici con lo scopo che sia proprio un medico l'utente finale.

Infine, si noti il valore di AUC di 0.889. Tale valore è molto elevato e permette di guardare alle probabilità stimate dai nodi foglia con alta confidenza. Le decisioni prese al terzo mese di terapia per l'HCV possono quindi essere prese osservando questo modello con una certa sicurezza. La ROC curve in Figura 9.8 è molto schiacciata verso l'alto: ciò consente di dire che le probabilità di essere LTR elevate stimate dal modello sono tutte legate a veri positivi.

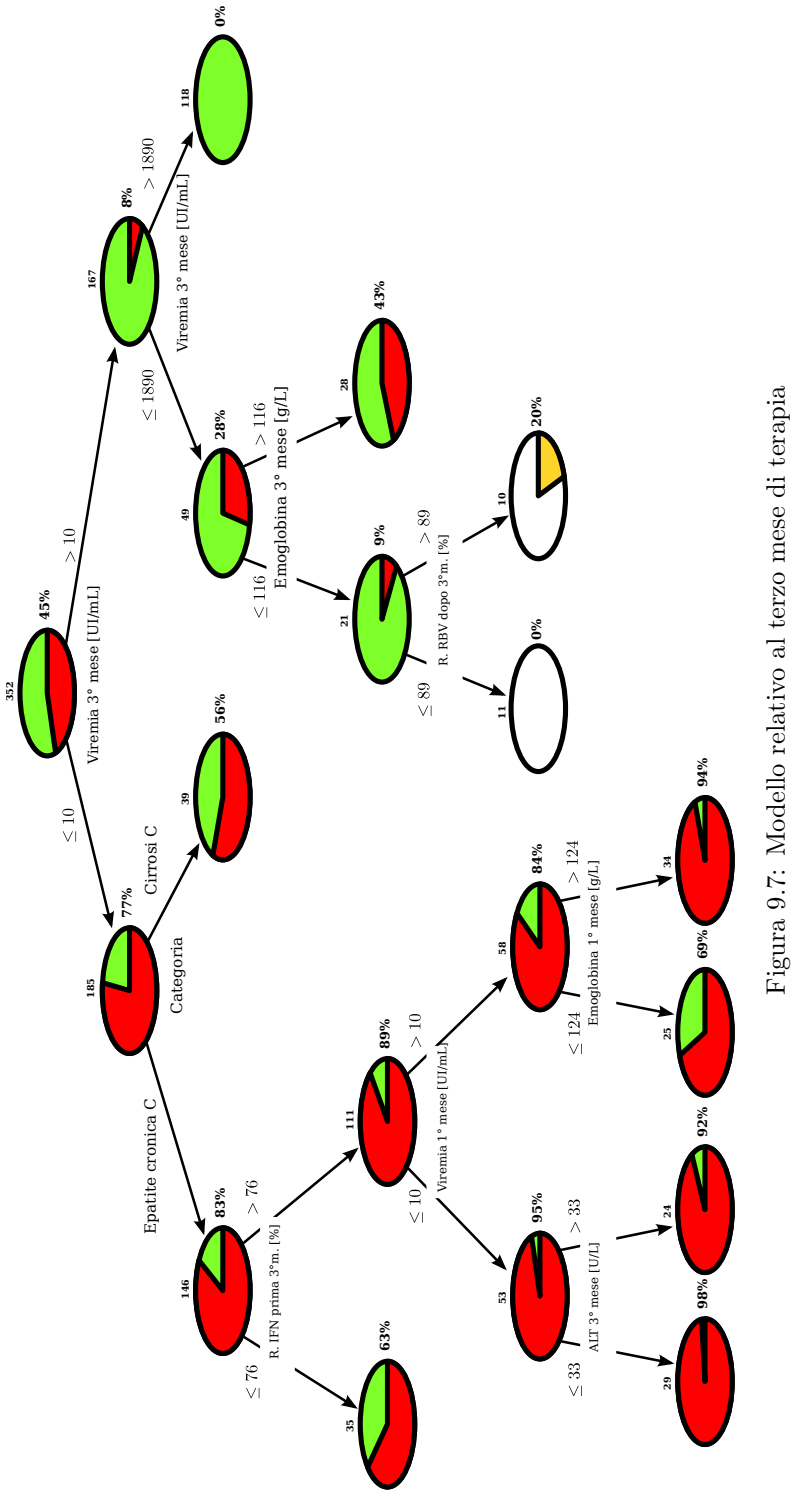


Figura 9.7: Modello relativo al terzo mese di terapia

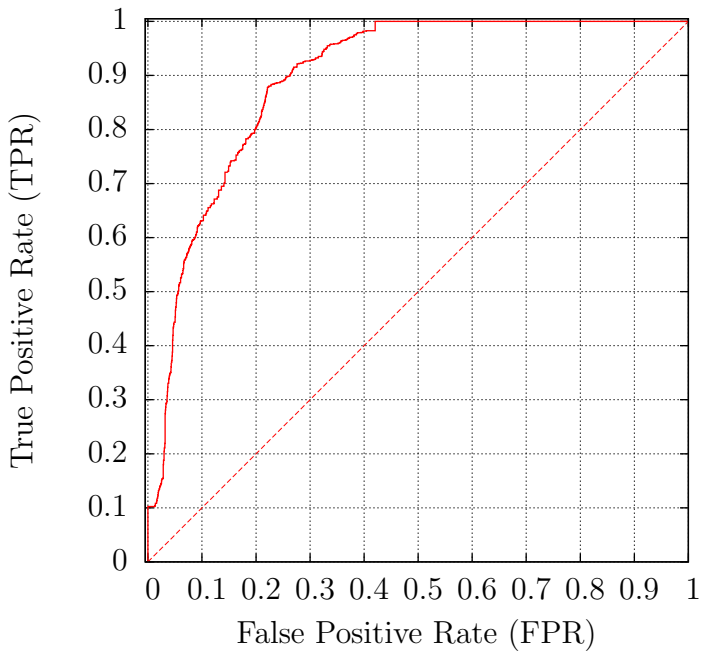


Figura 9.8: ROC curve disegnata per un modello costruito per il terzo mese di terapia con N_p e N_f rispettivamente 20 e 10

Capitolo 10

Conclusioni

La valutazione da parte di esperti del settore medico ha permesso di perfezionare le tecniche utilizzate per realizzare modelli di classificazione semplici, chiari e accurati. Dalla osservazione dei risultati ottenuti da tali modelli si evince che l'approccio adottato è efficace: essi possono quindi fornire un reale supporto alle decisioni cliniche. Il problema comunque molto complesso e necessita di essere studiato più a fondo. Questa tesi lascia aperte molteplici domande a cui si vorrebbe rispondere e su cui si deve concentrare l'attività di ricerca. Nel prossimo paragrafo vengono elencati i possibili sviluppi di questo lavoro. L'elevata interdisciplinarietà del progetto richiederà l'applicazione congiunta di competenze mediche e informatiche per poter proporre soluzioni mirate alla terapia attuale del paziente affetto da infezione HCV cronica.

10.1 Futuri ambiti di ricerca

Di seguito vengono elencati gli ambiti di ricerca in cui potrebbe essere ramificato questo lavoro di tesi:

- Seguendo le indicazioni dell'articolo di Provost sui PET [31], è possibile migliorare le stime di probabilità ottenute da tali modelli. Tramite confronti sulle AUC di test eseguiti su database dell'*UCI Machine Learning Repository* [35], viene asserito che è possibile ottenere stime più accurate introducendo la correzione di Laplace (si veda il Paragrafo 6.2) e il Bagging (si veda il Capitolo 4). La prima soluzione non modifica sostanzialmente la struttura dei PET e serve per ottenere stime di probabilità con valori meno estremi per campioni di record di piccole dimensioni. Tale approccio è stato seguito ma non ha permesso di migliorare le AUC rispetto a quelle dei test effettuati. La seconda soluzione purtroppo non permette di ottenere modelli a *scatola bianca*, cioè che possano rendere esplicito il percorso di scelta su è basata la classificazione (come nel caso di un PET). Come più volte discusso, questo è un requisito essenziale perché i classificatori possano avere una reale applicazione in ambito medico. Per questo si è tentato di usufruire dei benefici del Bagging mantenendo la trasparenza degli alberi decisionali con il metodo del Combine Multiple Model (trattato nel

Paragrafo 4.3). Anche questo tentativo non ha però sortito i risultati sperati: esso richiede infatti di essere perfezionato per essere applicato.

Si delinea quindi un obiettivo principalmente informatico a prosecuzione di questo lavoro: l'applicazione di tecniche che permettano di migliorare le stime di probabilità ottenute dai PET senza intaccare la trasparenza della loro rappresentazione;

- L'utilizzo dei rapporti di IFN e RBV, calcolati (come spiegato nel Paragrafo 7.3) come dose media di farmaco diviso dose standard, introduce l'implicita assunzione che la terapia per l'HCV genotipo 1 e 4 duri in tutti i casi 12 mesi, tranne per i soggetti che non raggiungono l'EVR il cui trattamento termina al terzo mese. I consigli proposti dai modelli presentati pertanto devono essere presi in considerazione con l'ipotesi di una durata di trattamento di 12 mesi e di una sospensione precoce in caso di non EVR. Questa approssimazione è stata necessaria per un primo approccio al problema. Gli sforzi internazionali si concentrano per ottenere quello che viene definito *Response-guided Therapy* o terapia guidata sulla risposta. Anche i modelli proposti mirano all'ottenimento di questo obiettivo: infatti a seconda della probabilità di essere LTR fornita dai PET ai vari istanti temporali è possibile scegliere se è il caso di terminare o meno una terapia inefficace. Soprattutto le informazioni fornite dagli alberi relativi al primo e al terzo mese di terapia sono fortemente legati all'obiettivo di Response-guided Therapy: a seconda di quello che accade entro tali periodi al paziente viene assegnata una probabilità di diversa di essere LTR e implicitamente proposta una durata di terapia diversa. L'obiettivo è di rendere esplicita questa proposta combinandola al consiglio sulle dosi più efficaci da assumere nel futuro che già i PET propongono;

- I medici epatologi che hanno studiato i modelli proposti in questa tesi sono sempre tentati, per acquisire una maggior quantità di informazioni, ad osservare più PET relativi allo stesso istante temporale ma ottenuti con parametri diversi. Questa è una buona metodica perché in alberi più compatti vengono proposti consigli la cui affidabilità è maggiore, mentre in alberi più sviluppati si evincono i legami non banali tra le variabili in gioco. Sarebbe allora molto interessante proporre un *software interattivo* che permetta al medico di reperire le informazioni necessarie dai vari modelli nel più breve tempo possibile. Anche l'opportunità di costruire l'albero in tempo reale potrebbe essere utile al medico per identificare la miglior terapia adatta a un nuovo paziente. Si immagini un software che, selezionando un nodo dell'albero relativo alla parte costruita con i parametri passati, proponga le eventuali alternative per la futura terapia. Tale software sarebbe di certo di interesse per i medici epatologi e potrebbe venire utilizzato per uniformare lo schema di trattamento in tutto il territorio Veneto e Nazionale.

Bibliografia

[1] M. G. Ghany, D. B. Strader, D. L. Thomas, and L. B. Seeff, "Diagnosis, management, and treatment of hepatitis c: An update," *Hepatology*, vol. 49, no. 4, pp. 1335–1374, April 2009.

[2] E. Sagnelli, T. Stroffolini, A. Mele, P. Almasio, N. Coppola, L. Ferrigno, C. Scolastico, M. Onofrio, M. Imparato, and P. Filippini, "The importance of hcv on the burden of chronic liver disease in italy: a multicenter prevalence study of 9,997 cases," *Journal of Medical Virology*, vol. 75, no. 4, pp. 522–527, April 2005.

[3] Q. L. Choo, G. Kuo, Weiner, L. R. Overby, D. W. Bradley, and M. Houghton, "Isolation of a cdna clone derived from a blood-borne non-a, non-b viral hepatitis genome." *Science (New York, N.Y.)*, vol. 244, no. 4902, pp. 359–62, April 1989.

[4] A. Alberti, L. Chemello, and L. Benvegnù, "Natural history of hepatitis c," *Journal of Hepatology*, vol. 31, no. Supplement 1, pp. 17 – 24, 1999.

[5] A. Alberti, F. Bonino, F. Bortolotti, M. Colombo, A. Craxì, A. Mele, and M. Rizzetto, "Trattamento della epatite da hcv: Raccomandazioni della associazione italiana per lo studio del fegato," Associazione Italiana per lo Studio del Fegato, Tech. Rep., 2004.

[6] L. Cavalletto, E. Bernardinello, G. Diodati, E. Raise, A. Gatta, and L. Chemello, "Terapia con peg-interferone e ribarina nell'epatite cronica c: costo-efficacia e farmacoutilizzazione nella comune pratica medica," *Farmeconomia e percorsi terapeutici*, vol. 9, pp. 173–181, 2008.

[7] E. Dieperink, M. Willenbring, and S. Ho, "Neuropsychiatric symptoms associated with hepatitis c and interferon alpha: A review." *Am J Psychiatry*, vol. 157, no. 6, pp. 867–76, 2000.

[8] S. Fagiuoli, V. G. Mirante, M. Pompili, S. Gianni, G. Leandro, G. L. Rapaccini, A. Gasbarrini, R. Naccarato, L. Pagliaro, M. Rizzetto, and G. Gasbarrini, "Liver transplantation: the italian experience," *Digestive and Liver Disease*, vol. 34, no. 9, pp. 640 – 648, 2002.

[9] J. M. Llovet, A. Burroughs, and J. Bruix, "Hepatocellular carcinoma," *The Lancet*, vol. 362, no. 9399, pp. 1907 – 1917, 2003.

[10] P. Colombatto, L. Civitano, F. Oliveri, B. Coco, P. Ciccorossi, D. Flichman, M. Campa, F. Bonino, and M. R. Brunetto, "Sustained response to interferon-ribavirin combination therapy predicted by a model of hepatitis c virus dynamics using both hcv rna and alanine aminotransferase," *Antiviral therapy*, vol. 8, pp. 519–530, 2003.

[11] Colombatto, P, Ciccorossi, P, Maina, A. M, Civitano, L, Oliveri, F, Coco, B, Romagnoli, V, Bonino, F, and M. R. Brunetto, "Early and accurate prediction of peg-ifns/ribavirin therapy outcome in the individual patient with chronic hepatitis c by modeling the dynamics of the infected cells," *Clin Pharmacol Ther*, vol. 84, pp. 212–215, 2008.

[12] H. Saito, H. Ebinuma, K. Ojiro, K. Wakabayashi, M. Inoue, S. Tada, and T. Hibi, "On-treatment predictions of success in peg-interferon/ribavirin treatment using a novel formula," *World J Gastroenterol*, vol. 16, pp. 89–97, 2010.

[13] M. Kurosaki *et al.*, "A predictive model of response to peginterferon ribavirin in chronic hepatitis c using classification and regression tree analysis," *Hepatology Research*, vol. 40, no. 3, pp. 251–260, March 2010.

[14] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to probability.* Athena Scientific, 2002.

[15] L. Soliani, "Manuale di statistica per la ricerca e la professione," Aprile 2005, disponibile online. [Online]. Available: http://www.dsa.unipr.it/soliani/soliani.html

[16] R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems, 5th Edition.* Benjamin/Cummings, 2007.

[17] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition).* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[18] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, 1995, pp. 1137–1145.

[19] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems.* Springer-Verlag, 2000, pp. 1–15.

[20] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[21] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[22] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] *Knowledge Acquisition from Examples Via Multiple Models.* Morgan Kaufmann, 1997.

[25] J. R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Chapman & Hall, New York, NY, 1984.

[27] L. Hyafil and R. L. Rivest, "Constructing optimal binary decision trees is np-complete," *Inf. Process. Lett.*, vol. 5, no. 1, pp. 15–17, 1976.

[28] S. K. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," *Data Min. Knowl. Discov.*, vol. 2, no. 4, pp. 345–389, 1998.

[29] C. I. Hovland and E. B. Hunt, "Programming a model of human concept formulation," *Managing Requirements Knowledge, International Workshop on*, vol. 0, p. 145, 1961.

[30] J. R. Quinlan and R. L. Rivest, "Inferring decision trees using the minimum description length principle," *Inf. Comput.*, vol. 80, no. 3, pp. 227–248, 1989.

[31] F. Provost and P. Domingos, *Tree Induction for Probability-Based Ranking.* Hingham, MA, USA: Kluwer Academic Publishers, 2003, vol. 52, no. 3.

[32] T. Fawcett, "An introduction to roc analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[33] K. Ishak, A. Baptista, L. Bianchi, F. Callea, J. D. Groote, F. Gudat, H. Denk, V. Desmet, G. Korb, R. N. MacSween, M. Phillips, B. G. Portmann, H. Poulsen, P. J. Scheuer, M. Schmid, and H. Thaler, "Histological grading and staging of chronic hepatitis," *Journal of Hepatology*, vol. 22, pp. 696–699, 1995.

[34] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, 2009.

[35] D. N. A. Asuncion, "Uci machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

Ringraziamenti

Desidero ringraziare *tutti* quelli che mi sono stati vicini in questo momento intenso. Nel particolare mia *mamma*, mio *papà*, mia sorella *Veronica* e mia *nonna Rosetta*.

Un ringraziamento speciale lo devo a *Filippo Bernardello* e *Stefano Mezzalira* per avermi aiutato ascoltandomi nei momenti in cui non avevo le idee chiare.

Ringrazio i Professori *Andrea Pietracaprina* e *Geppino Pucci* assieme alle Dottoresse *Luisa Cavalletto* e *Liliana Chemello* per la professionalità con cui mi hanno seguito in questo lavoro.

Per ultimo, ma non per importanza, voglio dire grazie a *tutti gli amici* esterni all'ambito accademico che hanno saputo farmi tornare alla realtà dopo lunghi periodi di studio.

Grazie,
Simone Romano