UNIVERSITA' DEGLI STUDI DI PADOVA

F A C O L T À D I S C I E N Z E S T A T I S T I C H E

CORSO DI LAUREA IN SCIENZE STATISTICHE ED EDCONOMICHE

TESI DI LAUREA

ANALISI DEI LIVELLI DI ESPRESSIONE GENICA PER LO STUDIO DELLE LEUCEMIE

Relatore: Chiar.ma Prof.
Monica Chiogna

Co-relatore: Dott.
Chiara Romualdi

Candidato
Martina Garlet

Anno Accademico 2002-2003

Alla mia famiglia

Indice

Capitolo 1 Introduzione ..... *- 1 -*

1.1 Obiettivi dell'elaborato ..... **- 2 -**

1.2 Alcune nozioni di biologia ..... **- 4 -**

1.3 La tecnologia dei DNA microarray ..... **- 7 -**

1.3.1 Preparazione dell'esperimento ..... - 8 -
1.3.2 Dall'immagine al dato ..... - 10 -
1.3.3 Normalizzazione ..... - 11 -
1.3.4 Filtrazione ..... - 13 -

1.4 I precedenti studi: una rassegna bibliografica ..... **- 14 -**

Capitolo 2 I dati ..... *- 25 -*

2.1 Il dataset ..... **- 25 -**

2.2 Imputazione dei valori mancanti ..... **- 27 -**

2.1.1 Tecniche di imputazione ..... - 27 -
2.1.2 Efficacia delle tecniche di imputazione ..... - 35 -
Appendice ..... **- 41 -**
A2.1 Valori mancanti ..... - 41 -
A2.2 Efficacia delle tecniche d'imputazione ..... - 42 -
A2.3 Routines R ..... - 44 -

Capitolo 3 Analisi esplorativa ..... *- 51 -*

3.1 Introduzione ..... **- 51 -**

3.2 Curve di Andrews ..... **- 51 -**

3.2.1 Curve di Andrews per il *dataset* completo ..... - 56 -
3.2.2 Curve di Andrews sul *dataset* ridotto ..... - 63 -

3.3 Analisi di raggruppamento sui geni ..... **- 67 -**

3.3.1 Analisi di raggruppamento: *dataset* completo ..... - 68 -
3.3.2 Analisi di raggruppamento: *dataset* ridotto ..... - 72 -
3.3.3 Analisi di raggruppamamento: un confronto ..... - 74 -
Appendice ..... **- 76 -**
A3.1 Curve di Andrews ..... - 76 -
A3.2 Analisi cluster ..... - 78 -
A3.3 Analisi cluster: dendrogrammi ..... - 87 -
A3.4 Confronto tra dataset completo e ridotto ..... - 96 -
A3.5 Curve di Andrews: liste di geni ..... - 100 -
A3.6 Routines R ..... - 106 -

Capitolo 4 Analisi discriminante ..... *- 113 -*

4.1 Introduzione ..... **- 113 -**

4.2 Alberi di classificazione ..... **- 113 -**

4.3 Analisi discriminante classica ..... **- 124 -**

4.3.1 Analisi basata sulle informazioni provenienti dagli alberi di classificazione con variabile risposta politomica. ..... - 127 -
4.3.2 Analisi basata sulle informazioni provenienti dagli alberi di classificazione per variabile risposta dicotomica. ..... - 128 -

4.4 Analisi discriminante logistica ..... - 133 -
Appendice ..... - 140 -
A4.1 Selezione dei geni nell'analisi discriminante ..... - 140 -
A4.2 Analisi discriminante classica e logistica sui medoidi ..... - 141 -

Capitolo 5 Metodi per il controllo della dimensionalità ..... - 143 -

5.1 Introduzione ..... - 143 -

5.2 Metodo " nearest shrunken centroid " ..... - 144 -

5.3 Confronto tra due metodi di selezione delle variabili ..... - 152 -

5.4 Regressione logistica penalizzata ..... - 161 -
5.4.1 Gruppo AML *vs* gruppo ALL ..... - 166 -
5.4.2 Gruppo ALL-B vs gruppo ALL-T ..... - 169 -
Appendice ..... - 174 -
A5.1 Metodo "shrunken centroids" ..... - 174 -
A5.2 Routines R ..... - 177 -

Capitolo 6 Considerazioni conclusive ..... - 189 -

Ringraziamenti ..... - 192 -

Riferimenti e bibliografia ..... - 194 -

Capitolo 1
Introduzione

Ormai da tempo, scienziati e studiosi concordano sulla teoria secondo la quale alcune malattie derivano da piccole alterazioni del codice genetico. Conseguentemente, a livello microscopico, ciò che distingue un individuo sano da uno malato, sono delle differenze nell'espressione dei geni ossia nel modo con cui essi sono utilizzati e nelle proteine a cui danno origine.

Un primo passo da compiere, muove pertanto verso la diagnosi: il problema è quello di caratterizzare l'anomalia genetica della cellula malata, ossia ciò che la differenzia rispetto a quella sana, in modo tale che, una volta noto il profilo genetico di un paziente, risulti possibile classificarlo come sano o affetto da malattia. Il passo successivo sarà diretto verso la cura di queste alterazioni mediante l'individuazione dei geni ai quali è imputabile l'insorgenza di una determinata patologia, o meglio l'individuazione dei geni che, in presenza di una determinata malattia, risultano alterati con maggior frequenza.

La tecnologia dei *DNA microarray* offre un valido strumento per l'analisi dell'espressione genica consentendo di comprendere gli stadi attraverso i quali una cellula sana diventa malata. Questo offre la possibilità di intervenire con terapie *ad hoc* o addirittura di assemblare molecole in grado di inibire il progredire della patologia, aumentando in tal modo l'efficacia terapeutica o ancora di bloccare quei geni che inducono la resistenza ai farmaci.

Cinque sono i principali obiettivi biologici che motivano l'applicazione della statistica nell'area dei *microarray*: (i) l'identificazione di geni differenzialmente espressi sotto diverse condizioni sperimentali o tra soggetti che presentano varie forme della stessa patologia; (ii) l'individuazione di gruppi di geni che con buona probabilità sono co-regolati; (iii) la

classificazione di campioni biologici (soggetto sano / soggetto malato); (iv) l'identificazione di geni detti marcatori[1] (*marker genes*) candidati come indicatori di un particolare un gruppo o fenotipo; (v) l'identificazione di nuove classi di una specifica patologia (es. il tumore).

1.1 Obiettivi dell'elaborato

Gli studi sul cancro costituiscono una delle maggiori aree di ricerca in campo medico. Un'accurata distinzione dei diversi tipi di tumori ha un'importanza fondamentale nel fornire trattamenti più mirati e rendere minima l'esposizione del paziente alla tossicità delle terapie. Fino a poco tempo fa, la classificazione delle varie forme di cancro ha sempre avuto basi morfologiche e cliniche; i metodi convenzionali però soffrono di diverse limitazioni soprattutto per quanto riguarda la capacità diagnostica. È stato recentemente suggerito che la differenziazione dei trattamenti in accordo con la differenziazione dei tipi di cancro, potrebbe avere un impatto positivo sull'efficacia della terapia, infatti le diverse classi tumorali presentano caratteristiche molecolari differenti oltre a distinti decorsi clinici.

Il livello di espressione genica contiene le chiave per affrontare problemi legati alla prevenzione e alla cura di alcune malattie, per comprendere i meccanismi di evoluzione biologica e per scoprire adeguati trattamenti farmacologici. Il recente avvento della tecnologia del DNA ha permesso di monitorare simultaneamente migliaia di geni, motivando lo sviluppo della classificazione delle forme di tumore con l'utilizzo di dati d'espressione genica. Sebbene questo tipo di approccio abbia mosso appena i primi passi, i

[1] I *geni marcatori* sono geni il cui valore di espressione è biologicamente utile per determinare la classe di tumore a cui appartiene il soggetto: in altre parole questi geni sono in grado di caratterizzare la classe tumorale.

risultati sono incoraggianti e sembrano dare spazio a svariate analisi ed interpretazioni.

Nel presente elaborato ci si propone lo studio delle leucemie soprattutto nell'ottica della classificazione. Lo scopo sarà pertanto quello di distinguere tra quattro tipi di leucemie sulla base dell'espressione genica dei pazienti che presentano questo tipo di affezione. L'interesse è in particolar modo rivolto all'individuazione dei geni che potrebbero essere responsabili dell'insorgere della patologia nonché della sua differenziazione nelle quattro forme che si vanno a presentare nel seguito.

Le leucemie sono malattie neoplastiche caratterizzate da una proliferazione tumorale dei tessuti linfoidi o mieloidi; nel primo caso si parla di malattie linfoproliferative, nel secondo di malattie mieloproliferative . In entrambi i casi, i tessuti presentano una diminuzione della differenziazione e un'alterazione importante dei meccanismi che controllano la morte cellulare programmata.

Possono presentarsi in forma acuta, se la proliferazione dei tessuti tumorali interessa cellule incapaci di differenziarsi e maturare completamente, oppure in forma cronica, se la proliferazione interessa cellule in grado di differenziarsi e maturare. Mentre si parla di un solo tipo di leucemie mieloidi acute (o AML *acute myeloid leukemia*), si possono distinguere tre tipi di leucemie linfoidi acute (ALL). A seconda dei linfociti che vanno a colpire, si parla di ALL-B (*acute linfoblastic leukemia B*) quando vengono intaccati i linfociti di tipo B, oppure di ALL-T (*acute linfoblastic leukemia T*) quando interessa i linfociti di tipo T; infine il terzo tipo, ALL-Tr (*acute linfoblastic leukemia traslocate*), comporta una particolare disomogeneità cromosomica dovuta alla traslocasione di alcuni frammenti di DNA all'interno dei cromosomi.

Le leucemie rappresentano la patologia tumorale che si osserva con maggior frequenza nella popolazione infantile (circa il 35% di tutte le

neoplasie). La forma più frequente è quella linfoblastica acuta (ALL) che si presenta nel 75% dei casi, mentre la forma mieloide acuta riguarda il 15% dei pazienti. Il restante 10% è invece costituito dalle leucemie mieloidi croniche e dalle mielodisplasie.

Le analisi che verranno proposte nel seguito sono state condotte utilizzando il *software* statistico R disponibile al sito `http//www.r-project.org.` In appendice a ciascun capitolo saranno riportate le *routines* di cui si è fatto uso nel corso dello stesso, sia quelle disponibili all'interno del *software* o delle specifiche librerie, sia quelle che si sono create *ex novo* (per queste ultime sarà fornito maggior dettaglio).

1.2 Alcune nozioni di biologia

Per comprendere meglio la trattazione della fase sperimentale, è opportuno fissare alcuni concetti base di biologia molecolare.

Le cellule sono le unità funzionali e strutturali biologiche di base. Sono separate dall'ambiente esterno da una membrana che, oltre a garantire l'integrità funzionale e strutturale della cellula, regola il passaggio delle sostanze dall'interno verso l'esterno e viceversa.

All'interno si trova il citoplasma, una soluzione acquosa concentrata attraversata e suddivisa da un elaborato sistema di membrane, il *reticolo endoplasmatico*, e contenente enzimi, ioni e molecole disciolte oltre ad un certo numero di organuli con funzioni specifiche. Tra questi organuli, rivestono particolare interesse i *ribosomi* che sono i siti in cui ha luogo l'assemblaggio e la sintesi proteica. Essi possono ricoprire il *reticolo endoplasmatico* oppure trovarsi liberi nel citoplasma. Oltre ai ribosomi, nel citoplasma hanno sede anche i *mitocondri,* in cui avvengono le reazioni chimiche che forniscono

energia per le attività cellulari, l'*apparato di Golgi,* dove sono immagazzinate le molecole sintetizzate nella cellula, i *lisosomi* e i *perossisomi,* che sono delle vescicole in cui le molecole vengono scomposte in elementi più semplici che possono essere usati dalla cellula oppure eliminati. Il citoplasma è inoltre fornito di un *citoscheletro,* che determina la forma della cellula, le consente di muoversi e fissa i suoi organuli.

Ma la struttura più grossa e più importante presente nella cellula è il *nucleo* che, interagendo con il citoplasma, aiuta a regolare le attività che si svolgono nella cellula. All'interno dell'involucro nucleare, formato da una doppia membrana, ha sede il *nucleolo*, il sito di formazione delle subunità ribosomiali nonché della *cromatina,* sostanza formata da un complesso di proteine e di DNA. Essa è la sostanza costitutiva dei cromosomi, è presente in tutto il nucleo e prende questo nome quando si trova in forma disciolta. Il DNA (acido deossiribonucleico) è una lunga molecola costituita da due filamenti avvolti l'uno sull'altro e uniti da dei ponti infinitesimali detti *ponti idrogeno*. I due filamenti sono composti da subunità ripetute di un gruppo fosfato e dello zucchero deossiribosio a cinque atomi di carbonio, mentre i ponti sono formati da una coppia di basi azotate. Uno zucchero deossiribosio, un gruppo fosfato e una base azotata costituiscono un *nucleotide*.

Esistono quattro tipi di basi: *adenina, timina, citosina, guanina* ed hanno la caratteristica di accoppiarsi sempre allo stesso modo, adenina con timina e citosina con guanina. Esse sono una sorta di alfabeto con il quale viene scandito il messaggio genetico: a seconda di come si presentano e si organizzano in triplette, si ha la formazione un particolare *gene*, che è per l'appunto un segmento di DNA in grado di trasmettere messaggi per la sintesi delle proteine ed altre sequenze regolative. Quando una molecola di DNA si duplica, i due filamenti si separano grazie alla rottura dei legami idrogeno e ciascuno, con le proprie basi azotate, funge da stampo per la formazione di un nuovo filamento complementare. È così che l'informazione ereditaria si

trasmette fedelmente da una cellula madre alla cellula figlia in quella che viene detta *duplicazione semiconservativa.*

La sequenza dei nucleotidi presenti nella molecola di DNA determina la sequenza degli amminoacidi ossia delle subunità necessarie per la sintesi proteica: una serie di tre nucleotidi (detta *codone*) codifica per un amminoacido. Il processo attraverso il quale il DNA viene tradotto in proteina consta di due fasi fondamentali: *trascrizione* e *traduzione*. Durante la prima fase, l'informazione viene *trascritta* da un filamento singolo di DNA in un filamento singolo di RNA detto messaggero o mRNA. L'RNA messaggero è una molecola del tutto simile al DNA solo che al posto della *timina* si trova un'altra base: l'*uracile*. Una volta trascritto, l'mRNA esce dal nucleo e si sposta sui ribosomi dove ha luogo la sintesi proteica o *traduzione*. I ribosomi sono costituiti da due subunità formate da RNA ribosomiale e proteine (o rRNA). A questo punto interviene RNA di trasporto o tRNA, una molecola che assume la forma di un trifoglio e provvede al trasporto degli amminoacidi. Il tRNA è munito di una tripletta di basi, detta *anticodone,* specifica per l'amminoacido che trasporta. Durante la sintesi, il tRNA mette in corrispondenza ciascuna tripletta di basi (*codone*) dell'mRNA con il suo anticodone, in modo che ogni molecola di tRNA apporti l'amminoacido specifico relativo al codone dell'mRNA a cui si attacca. In questo modo, in base alla sequenza dettata inizialmente dal DNA, le unità amminoacidiche vengono allineate una dopo l'altra andando ad assemblare la catena polipeptidica ossia la *proteina*.

Le mutazioni non sono altro che cambiamenti nella sequenza o nel numero di nucleotidi nell'acido nucleico della cellula, dovuti all'aggiunta, alla delezione o alla sostituzione di un nucleotide con un altro. Molte malattie genetiche sono il risultato della mancanza o inattività di enzimi o altre proteine. Queste a loro volta sono provocate da mutazioni nei geni che codificano per tali proteine.

Per comprendere la tecnica dei *microarray chip* è fondamentale notare che, nella fase di trascrizione, ciascuna cellula produce RNA solamente per quei geni (ossia per quei segmenti di DNA) che sono attivi in quel momento; pertanto, un modo per indagare quali sono i geni attivi e quali quelli inattivi in un determinato istante, sarà quello di analizzare l'RNA prodotto dalla cellula, ed è da questo punto che parte l'intuizione della *DNA microarray tecnology*.

1.3 La tecnologia dei DNA *microarray*

L'idea base della *microarray tecnology* è di misurare simultaneamente il livello di espressione relativo di centinaia di geni di una particolare popolazione di cellule o di un tessuto. Due sono i concetti chiave alla base di questo processo di misura: la *retro trascrizione* e l'*ibridizzazione*.

La retro trascrizione è il processo inverso rispetto alla trascrizione. Nella trascrizione, un filamento singolo di DNA viene usato come stampo per la costruzione di una molecola di RNA. Durante la retro trascrizione, avviene il contrario: una molecola di mRNA viene in questo caso *retro-trascritta* in DNA complementare o cDNA. Si noti che l'mRNA di una cellula può essere isolato sperimentalmente e retro trascritto andando così a costituire la libreria di cDNA della cellula. Inoltre il cDNA rappresenta una molecola molto più stabile dell'mRNA e quindi più adatta ad essere studiata e manipolata.

L'ibridizzazione è invece un processo di appaiamento delle due eliche singole di DNA o RNA. Esse si separano ad una temperatura caratteristica di 65^{o} C. Se la temperatura viene fatta scendere e mantenuta al di sotto di quella di separazione, le due eliche tornano ad unirsi con la loro controparte. Tale processo è basato sul principio di appaiamento della basi che consente l'unione, ossia l'ibridizzazione, solo di segmenti di DNA o RNA tra loro complementari.

La *microarray tecnology* viene usata per misurare il livello relativo di espressione dei geni in un particolare tessuto ibridizzando il cDNA retro trascritto dall'mRNA di una cellula con delle sequenze di cDNA sintetico create in laboratorio e depositate in un *microchip* (o *array*). Il cDNA a contatto con il chip viene chiamato *sonda* (*probes*) mentre il cDNA derivante dalla retro trascrizione dell'mRNA cellulare prende il nome di *target*.

1.3.1 Preparazione dell'esperimento

La realizzazione del micro*array* consta di due fasi: la preparazione del *microchip* e quella del *target*.
Ad un *vetrino* (*microchip*) si fissano delle sonde (*probes*) costituite da segmenti di cDNA sintetico che riproducono i geni che in qualche modo sono notoriamente correlati con la patologia oggetto di studio. Per questo scopo esistono speciali *robot* in grado di dispensare goccioline dell'ordine di nanolitri attraverso tubi e punte eccezionalmente sottili.

Per preparare il *target*, si estrae l'mRNA totale prodotto dai due tipi di cellule in analisi. Per mezzo di una reazione biochimica l'mRNA viene retro trascritto dando luogo al cDNA che, come ricordato precedentemente, rappresenta una molecola più stabile dell'mRNA. Durante questa fase, nella catena di cDNA di ciascun gene, vengono introdotte particolari molecole dette *recetttori* in grado di legarsi a sostanze fluorescenti. Successivamente il cDNA dei due tipi di cellule viene "etichettato" con due colori (rosso e verde) mediante dei marcatori fluorescenti che vanno a legarsi ai recettori: Cy3 (verde) per le cellule sane e Cy5 (rosso) per quelle malate. Infine il cDNA delle due cellule viene mescolato e depositato sull'*array* affinché possa ibridizzare con le sonde. Durante l'ibridizzazione, i segmenti di cDNA *target* riconoscono le sonde complementari e si legano ad esse.

Una volta completata l'ibridizzazione, il *microchip* viene lavato e successivamente eccitato con un laser affinché i marcatori fluorescenti

emettano un segnale luminoso. Uno speciale scanner legge l'*array* illuminando ciascuno *spot* (ossia ciascun puntino che rappresenta di un singolo gene) e misurando la fluorescenza emessa per ciascun colore separatamente, in modo da fornire una misura della quantità relativa di mRNA prodotto da ciascun gene nei due tipi di cellula.

L'intensità degli *spot* verdi misura la quantità relativa di cDNA contrassegnato con Cy3, e quindi mRNA prodotto dalle cellule sane; mentre quella degli *spot* rossi misura la quantità relativa di cDNA segnato con Cy5, e quindi di mRNA prodotto dalle cellule malate. Queste misure forniscono informazioni sul livello relativo d'espressione di ciascun gene nelle due cellule. Le due immagini monocromatiche (rossa e verde) vengono poi sovrapposte in modo da fornire una visione d'insieme: ciascuno *spot* corrisponde ad un gene ed il colore alla sua condizione nella cellula malata e in quella sana. Così il rosso corrisponde ad un gene molto attivo nella cellula malata e inattivo in quella sana, il nero ad un gene inattivo in entrambe le cellule, il giallo ad un gene ugualmente attivo nei due tipi di cellule, ed infine il verde ad un gene attivo nella cellula sana e inattivo in quella malata.

È necessario che queste misure vengano aggiustate per considerare un disturbo di fondo causato ad esempio dall'alta concentrazione di sale e detergente durante l'ibridizzazione o la contaminazione del *target* o da altri problemi che si presentano durante l'esperimento. Delle apposite scale di misura associano poi a ciascuna tonalità di colore un rapporto tra l'attività dello stesso gene nei due tipi di cellula.

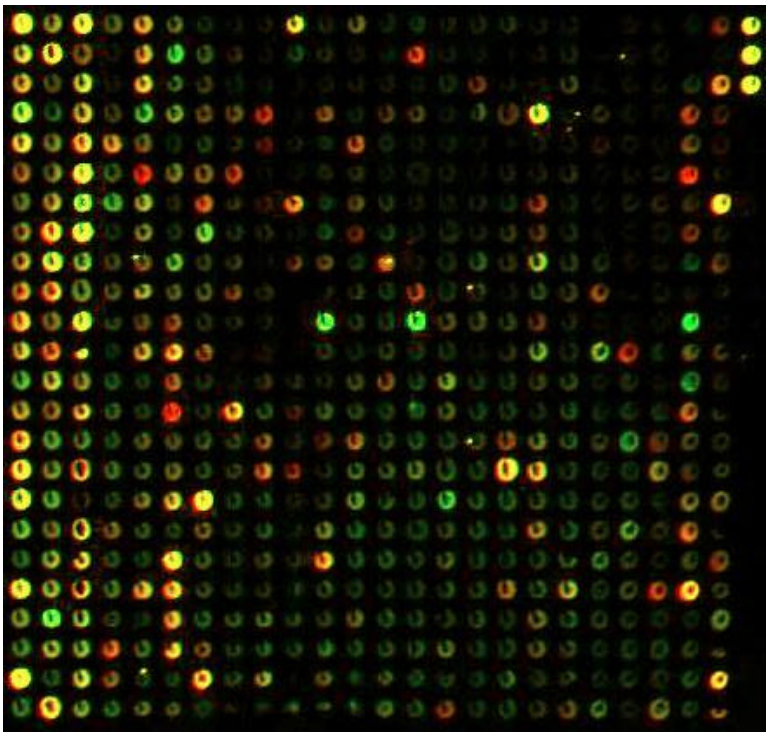

Figura 1.1 Immagine ottenuta con la tecnologia dei *microarray*. L'intensità del colte di ciascun punto (*spot*) codifica l'informazione sullo specifico gene del campione analizzato.

1.3.2 Dall'immagine al dato

L'ibridizzazione del *target* alle sonde determina una reazione chimica che viene catturata in un'immagine digitale da uno scanner laser. Il passo successivo è quello di tradurre l'intensità del segnale luminoso emesso da ciascun gene, in un coefficiente numerico. S'intuisce pertanto l'importanza della qualità dell'immagine ai fini di un'accurata interpretazione dei dati. I passi principali nell'analisi delle immagini prodotte da *cDNA microarray* sono: grigliatura (*gridding*), segmentazione ed estrazione di intensità.

La *grigliatura* ritrova nell'immagine, le posizioni degli *spot* che corrispondono alle sonde. Essendo nota la posizione degli *spot* nel micro*array*, questa operazione non risulta particolarmente complessa, sebbene si renda

necessaria la stima di alcuni parametri per tener conto ad esempio di shifts o rotazioni del micro*array* nell'immagine o di piccole traslazioni degli *spot*.

L'*estrazione d'intensità* calcola invece l'intensità della fluorescenza rossa e verde, l'intensità del *backgroud* ed alcune misure di qualità.

La *segmentazione* consiste infine nel separare il segnale emesso dai marcatori fluorescenti (*foreground*) rispetto la disturbo di fondo (*backgrund*), in modo da isolare le quantità di interesse.
Può succedere che questa correzione abbia l'effetto indesiderato di introdurre valori negativi (questo succede quando l'intensità di *background* è più forte di quella di *foreground*). In tal caso questi *spot* vengono trascurati oppure il loro segnale è sostituito con un valore arbitrariamente piccolo e positivo.

1.3.3 Normalizzazione

Al fine di rendere comparabili i risultati ottenuti su *array* diversi o anche all'interno dello stesso *array*, è necessaria la rimozione di alcune distorsioni sistematiche introdotte nella fase di preparazione dell'*array* stesso, di esecuzione dell'esperimento, nonché nel processo di ibridizzazione e nella scansione con il laser. La procedura di normalizzazione si riferisce proprio al trattamento statistico dei dati finalizzato alla rimozione di tali effetti distorsivi.

Ad esempio, un diffuso problema nell'interpretazione dei dati derivanti da micro*array*, noto come *dye-effect*, è la diversa intensità di fluorescenza dei due marcatori Cy3 (verde) e Cy5 (rosso), cosicché l'emissione di fluorescenza del verde è sistematicamente meno intensa di quella del rosso. Il modo più immediato per rimuovere questo tipo di distorsione, sarebbe quello di ripetere due volte l'esperimento scambiando l'assegnazione dei marcatori tra i due *target*, cosa che però renderebbe la tecnica ancora più dispendiosa. Un'altra fonte di distorsione, nota come *print-tip effect*, è dovuta alla diversa quantità di materiale genetico (*probe*) depositata sul vetrino a causa di microscopiche differenze nella conformazione delle puntine del *robor* che stampa l'*array*.

Infine un terzo tipo di alterazione, l'*array-effect*, può derivare da differenze di intensità tra un *array* e l'altro legate a diverse condizioni di preparazione (usura delle puntine, qualità di conservazione e quantità dei reagenti), estrazione (differenti quantità di mRNA usate per creare il *target* o quantità di marcatore fluorescente), ibridizzazione (cross-ibridation) e scansione (bilanciamenti dei laser, diversi parametri di scansione).

Ai problemi sopra esposti si cerca di dare soluzione mediante il processo di normalizzazione. La normalizzazione prevede che si calcolino fattori di standardizzazione per ciascuno dei tre effetti sopra menzionati. Si tratta di sottrarre al segnale una media generale di *array*, la differenza tra la medie degli *spot* stampati da ciascun *print-tip* e la media generale ed infine la differenza tra la media delle intensità con fluorescenza rossa e verde.

Anzitutto il ricercatore deve scegliere quali geni usare nel processo di standardizzazione. Questa decisione è influenzata da alcune considerazioni come la proporzione attesa di geni differenzialmente espressi e la possibilità di controllare le sequenze di DNA. Tre sono gli approcci principali. Il primo si fonda sull'assunzione che solo una piccola parte dei geni sia differenzialmente espressa. I restanti geni hanno pertanto un livello d'espressione costante e possono essere usati come indicatori dell'intensità relativa dei due colori. In altri termini, quasi tutti i geni dell'*array* possono essere usati per la normalizzazione quando si può ragionevolmente assumere che solo una piccola porzione di essi vari significativamente la propria espressione da un campione all'altro, oppure che esista simmetria nei livelli di espressione dei geni sovra e sotto espressi. In pratica è però molto difficile trovare un gruppo di *spot* con un segnale costante su cui tarare un fattore di correzione. Si preferisce quindi, quando il numero di geni differenzialmente espresso è limitato rispetto al numero totale di geni indagati, usare tutti gli *spot* dell'*array* nel processo di normalizzazione dei dati. Il secondo approccio si basa sull'assunto che la proporzione di geni differenzialmente espressi sia alta e quindi suggerisce l'uso della restante porzione (*housekeeping genes*) che si crede abbia un livello di

espressione costante nelle due condizioni. Questa piccola porzione di geni però, oltre ad essere difficilmente identificabile, spesso risulta poco rappresentativa rispetto ai geni di interesse essendo costituita per lo più da geni con un alto livello di espressione. Il terzo approccio necessita dell'appoggio del laboratorio e prevede di realizzare un *microarray* per un solo campione di mRNA (prelevato da un'unica cellula) diviso in due porzioni uguali, ciascuna marcata con colori differenti. Trattandosi dello stesso campione di materiale genetico, in seguito all'ibridizzazione si dovrebbe avere la stessa intensità degli *spot* per il rosso e per il verde: eventuali differenze possono essere usate come fattore di normalizzazione.

1.3.4 Filtrazione

Un altro trattamento dei dati preliminare all'analisi è la cosiddetta *filtrazione*. Essa è finalizzata alla riduzione della variabilità e della dimensionalità dei dati. Il primo obiettivo vene raggiunto rimuovendo quei geni le cui misure non sono sufficientemente accurate, il secondo con l'eliminazione dei geni che non risultano abbastanza distinti. Tipicamente vengono rimossi i geni che presentano un livello di espressione molto piccolo o negativo (prima o dopo la normalizzazione). In pratica, tutti gli *spot* in cui la differenza tra l'intensità di *foreground* e quella di *background* non supera un valore soglia di 1.4-*fold* vengono eliminati o sostituiti con un valore piccolo arbitrario. Questa procedura è giustificata dall'evidenza empirica che livelli di espressione più piccoli di 10 sono solitamente frutto di errori di misura. Nonostante ciò, generalmente i ricercatori tendono ad adottare criteri meno stringenti per evitare di trascurare significative porzioni di dati.

Si noti che qualsiasi operazione di filtrazione introduce arbitrarietà nella scelta delle soglie che determinano se un valore è troppo grande o troppo piccolo oppure se la variabilità delle misure è troppo elevata.

1.4 I precedenti studi: una rassegna bibliografica

In letteratura sono stati applicati diversi metodi di classificazione per l'analisi e la distinzione delle forme di cancro, ma esistono alcuni problemi che rendono questo compito tutt'altro che banale. I dati di espressione genica sono infatti diversi da quelli con cui è abituato a trattare normalmente lo statistico. A questo proposito, Gordon K. Smith *et al.* (2002) rif [1] hanno stilato una rassegna dei problemi statistici e non che possono presentarsi nell'analisi di espressioni geniche dalla fase sperimentale fino alle diverse tecniche di trattamento del dato mettendo in luce eventuali ripercussioni sui risultati.

Il primo problema che pone il *dataset* è la sua grande dimensionalità (qualche migliaia di geni) a cui si contrappongono campioni molto limitati (di solito meno di un centinaio di unità). In letteratura si fa riferimento a questo problema con l'espressione "*large p and small n*". Oltre a comportare spesso tempi di elaborazione piuttosto lunghi, questa caratteristica espone al rischio di sovraparametrizzazione dei modelli tanto per l'alta dimensionalità quanto per la limitata numerosità campionaria. Se si immagina infatti di mappare le osservazioni nello spazio degli attributi, il campione può essere visto come una serie di punti piuttosto radi in un spazio multidimensionale molto ampio. In secondo luogo, la maggior parte dei geni presenti nel *dataset* sono irrilevanti ai fini della classificazione e costituiscono *rumore* che interferisce con il potere discriminante degli altri geni. Questo accresce non solo i tempi di computazione, ma anche le difficoltà della classificazione. È evidente che i metodi di classificazione esistenti non sono stati concepiti per essere applicati su questo tipo di dati. Alcuni ricercatori propongono il raggruppamento dei geni in classi omogenee come operazione preliminare alla classificazione dei soggetti, in quanto questo può aiutare a ridurre la dimensionalità, i tempi di calcolo e ad eliminare i geni irrilevanti che comportano una minor accuratezza nella classificazione. Una terza questione riguarda la natura stessa dei dati che sono caratterizzati dalla massiccia presenza di rumore che può essere distinto

in rumore *biologico* e *tecnico*; il primo si riferisce al rumore introdotto dai geni non rilevanti per la classificazione, il secondo comprende invece le componenti di disturbo introdotte nei vari stadi di preparazione del *dataset*.

Ma oltre a problemi di rilievo prettamente statistico, ci sono diverse questioni derivanti dal contesto biologico e dall'importanza dei risultati in campo medico. Una di queste riguarda la corrispondenza tra rilevanza biologica e statistica di un gene come classificatore: l'accuratezza e l'efficienza della regola di classificazione sono concetti importanti dal punto di vista statistico, ma non devono diventare l'unico obiettivo dell'analisi dei dati derivanti da *microarray*. La rilevanza biologica è un altro criterio da tenere in forte considerazione, in quanto ogni informazione rivelata durante il processo di analisi può essere utile per la scoperta delle funzioni specifiche di un certo gene, per la determinazione di gruppi di geni che concorrono allo sviluppo di cellule o tessuti cancerosi, per la scoperta di interazioni tra i geni o per altri studi biologici come l'individuazione dei geni marcatori. Infine esiste la questione della cosiddetta *contaminazione del campione* che va ad inserirsi nella problematica riguardante il conflitto tra rilevanza biologica e statistica. Usualmente infatti tessuti normali e cancerosi sono composti da cellule differenti: il tessuto tumorale è ricco di cellule epiteliali, mentre il tessuto normale è formato da una grossa porzione di cellule muscolari. Questo può condurre alla selezione di geni che hanno diversi valori di espressione nei due tessuti, ma tale differenza è imputabile diversa composizione dei tessuti stessi e non alla distinzione canceroso / normale. L'usuale errore nella selezione dei geni o nella classificazione è dunque quello di dividere il campione sulla base di quei geni che distinguono la composizione cellulare anziché di quelli correlati con il cancro. Questo ha come risultato una corretta classificazione ma non fornisce elementi biologici rilevanti circa il riconoscimento di geni correlati con la patologia.

La letteratura, basandosi su tecnologie piuttosto recenti, non è molto datata e tende a svilupparsi di pari passo con le scoperte in ambito biologico,

entrambe le discipline infatti traggono reciprocamente beneficio e materiale di studio l'una dallo sviluppo dell'altra e sono fortemente intrecciate in questo processo di crescita.

Nel 2001 Rocke e Durbin rif [2] introducono un modello di misura per gli errori nei dati da *microarray* come funzione del livello di espressione dei geni. Questo modello, assieme ad altri metodi di analisi e trasformazione dei dati, permette una più precisa comparazione tra le espressioni dei geni oltre a fornire una guida per l'analisi del segnale di *background*, la determinazione di intervalli di confidenza ed il trattamento preliminare dei dati.

Nel 1999 Lausen rif [3] si concentra sulle misure di distanza allineando sequenze di dati (per esempio la sequenza del DNA) secondo diversi criteri, propone poi un grafico (*dot-matrix plot*) come possibile *test* sulla bontà dell'allineamento; inoltre poiché l'assunto di identica distribuzione lungo le sequenze non è sempre appropriato, discute un metodo che permette l'esplorazione combinata della distribuzioni dei geni in diverse posizioni; nello stesso anno Golub *et al* rif [4] applicano su un campione di dati derivanti da leucemie di tipo acuto l'analisi *cluster* e l'analisi discriminante proponendo il *weighted gene voting scheme*. Jean Clavarie rif [5] rivede invece l'approccio teorico e computazionale usato fino ad allora per identificare i geni differenzialmente espressi (attraverso diversi tipi di cellule , diversi stadi di sviluppo, condizioni patologiche ecc.), per selezionare i geni co-regolati attraverso un insieme di condizioni e per creare *cluster* di geni che raggruppino in modo coerente caratteristiche di espressione simili. Nell'ottobre dello stesso anno Golub *et al* rif [6] applicano due procedure di classificazione (*class discovery* e *class prediction*) per distinguere diversi tipi di cancro su un *dataset* di leucemie acute. La *class discovery* è in grado di distinguere tra leucemie ALL e AML senza l'introduzione della variabile risposta, mentre la *class prediction* prevede in modo preciso la classe di appartenenza di una nuova osservazione. I risultati dimostrano inoltre che è possibile classificare le forme cancerose basandosi solo sul monitoraggio dell'espressione genica e

suggeriscono una strategia generale per la scoperta e la previsione di classi tumorali anche per altri tipi di cancro indipendentemente delle conoscenze biologiche che si hanno a priori. Platt (1999) rif [7] mette a punto la "*sequential minimal optimization*" che permette l'implementazione delle *support vector machines* (SVM) per affrontare problemi di classificazione che coinvolgono grandi *dataset* come quelli riguardanti le espressioni geniche. L'anno successivo Brown *et al* rif [8] testano diverse SVM usando varie misure di somiglianza su dati da *microarray* trovando che le SVM garantiscono prestazioni migliori rispetto ad altre tecniche nel riconoscere con successo geni coinvolti nelle comuni funzioni biologiche, mentre Ben Dor (2000) *et al* rif [9] descrivono una applicazione di SVM con nuclei lineare e quadratico che ha classificato con successo tessuti normali e tumorali del colon. Alizadeh *et al* sempre nel 2000 rif [10] analizzano *dataset* sul cancro ed usano tecniche di raggruppamento gerarchico per studiare l'espressione genica nelle tre prevalenti forme di tumore linfoide che colpisce gli adulti. Essi trovano l'analisi di raggruppamento su livelli di espressione genica un valido strumento essendo in grado di riconoscere e separare sia processi biologici che diversi tipi di cellule. Golub et al (2000) rif [11] partendo da un campione di 6817 geni e 38 pazienti creano una regola per distinguere tra leucemie ALL ed AML formando due *cluster* in cui raggruppano geni simili. Successivamente eseguono un'analisi *nearest neighbor* con cui identificano 1100 geni il cui livello cambia in relazione all'appartenenza al gruppo ALL o AML. Poi scelgono un insieme di 50 geni per costruire una regola di classificazione che assegna 29/43 pazienti correttamente; infine usano la tecnica SOM (*self organizing map*) al posto del *clustering* iniziale producono solo 4/38 errori ed applicando la stessa tecnica con un sottoinsieme di 20 geni ottengono ancora 4/38 errori. Veer *et al* (2000) rif [12] studiano un *dataset* di 78 pazienti con cancro al seno. Partendo da 5000 geni si restringono a 231 esaminando il coefficiente di correlazione di ciascun gene con il risultato della prognosi. Sempre nello stesso anno Ben-Dor *et al* rif [13] verificano che mentre le

support vector machine (SVM) hanno maggiore accuratezza sui dati da leucemie e i metodi basati sul *cluster*ing funzionano meglio su un *dataset* di tumori del colon, il metodo *nearest neighbor* dà buoni risultati in entrambi i casi. Keller *et al* (2000) rif [14] comparano il metodo bayesiano semplice con il metodo *weighted voting*: il primo sembra dare risultati migliori per *dataset* riguardanti leucemie e tumori alle ovaie, il secondo per *dataset* su tumori al colon. Nell'agosto 2000 Keller *et al* rif [15] presentano il metodo bayesiano semplice per la classificazione di tipi di tessuto con dati da *microarray* usando una tecnica basata sulla massima verosimiglianza per selezionare il gruppo di geni più utili a fini classificatori. Applicando questa tecnica ad un *dataset* con due tipi di tessuti riscontrano un'eccellente accuratezza e fanno notare che è facilmente estendibile ad un classificazione con più di due classi fornendo ottimi risultati se applicato ad un *dataset* con tre tipi di tessuto. Gen Hori *et al* (2000) rif [16] dimostrano l'applicazione del metodo ICA (indipendent *component analysis*) che è in grado di classificare un vasto insieme di dati di espressione genica in gruppi significativi dal punto di vista biologico. In particolare dimostrano che geni la cui espressione è campionata a diversi istanti temporali possono essere classificati in gruppi differenti e che questi gruppi hanno una buona somiglianza con quelli che si determinano solo sulla base delle conoscenze biologiche; questo suggerisce anche che il metodo ICA può essere uno strumento potente per la scoperta di ignote funzioni biologiche dei geni. L'anno successivo Zhang *et al* rif [17] e Kerr *et al* rif [18] usano il metodo *bootstrap* per valutare la qualità dell'analisi di raggruppamento i primi assumendo che i livelli di espressione hanno distribuzione normale, i secondo usando un modello ANOVA per generare i campioni *bootstrapped*. Steinwart (2001) rif [19] propone delle trasformate che rendono le osservazioni più facilmente separabili. Khan *et al* (2000) rif [20] si occupano delle *linear neural networks* per analizzare dati derivanti da pazienti affetti da quattro tipi di cancro. Viene usata l'analisi delle componenti principali per ridurre i 2308 geni in 10 componenti. I *networks* vengono poi analizzati per produrre una lista di

96 geni che sono in grado di classificare in modo corretto i 63 pazienti. Dehanasekaran *et al* (2000) rif [21] riducono un *dataset* di 9984 geni a 34 che classificano senza errori 53 soggetti in tre classi. Nel maggio 2001 David B., Allison *et al* rif [22] sviluppano una sequenza di procedure che comprendono modelli misti e inferenza *bootstrap* per affrontare i problemi (come *large p and small n*) che sorgono nel trattamento di dati che comprendono l'espressione di centinaia di geni, illustrando queste tecniche con *dataset* disponibili in rete. I due assunti di cui si servono nel loro lavoro sono: (i) ciascun gene ha media e varianza finite; (ii) le misure di ciascun gene sono sufficientemente affidabili. Nel luglio dello stesso anno Lorenz Wernisch rif [23] propone una rassegna dei principali metodi di trattamento dei dati da *microarray* fornendo alcuni interessanti dettagli spesso omessi da autori di altri articoli. Tibshirani *et al* (2001) rif [24] propongono una quantità per la stima del numero di gruppi (*cluster*) in un *dataset*: l'idea chiave è di vedere il *clustering* come un problema di classificazione *supervised* in cui bisogna stimare la vera classe di appartenenza. Questa quantità suggerisce quanti gruppi devono essere formati e quanto affidabile è la previsione. Nel corso dell'analisi sviluppano anche una nuova nozione di distorsione e di varianza per dati senza la variabile risposta (*unlabelled*). Queste tecniche vengono poi applicate ad un *dataset* sul cancro al seno. Infine vengono stabilite e studiate alcune misure di consistenza. Nello stesso anno Terese Biedl *et al* rif [25] elaborano un modo per rappresentare dati di espressioni geniche per rendere evidenti somiglianze e/o differenze cercando un ordinamento lineare dei geni tale che geni con profili di espressione simili vengano affiancati. Poiché non è facile trovare l'ordinamento migliore, propongono di cercarlo a seguito di una procedura di raggruppamento (*clustering*) e dimostrano che questa procedura consente di trovare con efficienza tale ordinamento rispettando le divisioni suggerite dall'analisi di raggruppamento. Katherine Pollare *et al* (2001) rif [26] identificano *pattern* complessi di dati considerando nelle procedure di *clustering* geni e soggetti simultaneamente. Propongono una classificazione a

due vie. Una stima del parametro che regola il raggruppamento si ottiene come funzione della vera distribuzione che genera i dati, e la sua stima si ha applicando questa funzione alla distribuzione empirica. I risultati vengono presentati ed illustrati su un *dataset* disponibile in rete. Sunduz Keles *et al* rif [27] concentrandosi su una singola condizione sperimentale, utilizzano i dati di espressione genica per identificare sequenze di geni che sono attivi sotto le stesse condizioni sperimentali. Fanno uso del modello lineare per modellare l'interazione a due vie nell'espressione genica; le variabili più rilevanti vengono selezionate con un metodi denominati *stepwise selection* di Monte Carlo. Applicando questa tecnica ad un *dataset* di 800 coppie di basi (questa volta i dati provengono da *oligonucleotide microarray*) vengono identificati con successo gli elementi regolatori che sono notoriamente attivi sotto le condizioni sperimentali in questione ed altre sequenze significative di oligonucleotidi che possono rappresentare nuovi elementi regolatori.

Nel febbraio 2002 Jie Liang Seman Kachalo rif [28] applica il metodo delle silhouette di P.J. Rousseauw per valutare la bontà del *cluster*ing e fa uso della teoria di Chervonenkis nota come *VC theory* che provvede alla ricerca di un iperpiano che massimizza la distanza tra le due osservazioni più vicine che appartengono a classi diverse. Parmigiani *et al* (2002) rif [29] propongono un modello utile per organizzare lo sviluppo di strumenti esplorativi per la classificazione. Questa struttura fa uso di variabili latenti per fornire una definizione statistica di *gene differenzialmente espresso* e di *profilo molecolare*. Ciò genera anche una misura di somiglianza per il l'analisi di raggruppamento tradizionale e dà luogo ad affermazioni probabilistiche circa l'assegnazione di uno certo tipo di tumore ad uno specifico profilo molecolare. Chris Fraley e Adrian E. Raftlery (2002) rif [30] rivedono una metodologia generale dell'analisi di raggruppamento che fornisce un approccio statistico a problemi come il numero di *cluster* da formare, il trattamento dei dati anomali (*outlyers*), il tipo di legame da usare ecc. Dimostrano anche che questa metodologia può essere utile nei problemi di analisi multivariata come l'analisi

discriminante o la stima di densità multivariate. Infine ne discutono le limitazioni e presentano i recenti sviluppi per dati non normali e *datasets* con grande dimensionalità. Darlene Goldsten *et al* rif [31] illustrano i problemi che possono insorgere nell'analisi *cluster* di *datasets* che raccolgono profili genetici, come il possibile alto grado di dipendenza dei risultati dalla scelta dell'algoritmo, dalla scelta dei geni o del campione. Propongono anche l'uso della regressione di Cox come strumento per l'identificazione dell'influenza dei geni sulla *genes influencing survival*. Kathleen Kerr *et al* (2002) rif [32] usano l'analisi della varianza (ANOVA) per il trattamento di dati da *microarray* per confrontare le espressioni geniche di soggetti trattati con farmaci e non trattati illustrando come questa tecnica possa essere applicata a dati che provengono da replicazioni di esperimenti, mentre metodi *ratio-based approch* non forniscono strutture per il trattamento di replicazioni che però sono chiaramente desiderabili quando si ha a che fare con dati che sono affetti da così tanto rumore. Descrivono anche i problemi che sorgono nell'analisi di dati da *microarray* come la scelta di un modello, l'inferenza e il *data scaling*. Tibshirani *et al* rif [33] propongono un metodo chiamato *cluster scoring* per *unsupervised learning*. Questo metodo generalizza quelli di raggruppamento dei geni sulla base della loro correlazione con una variabile risposta. Iniziando con un'analisi di raggruppamento sui geni, calcolano poi un punteggio per il singolo gene e per la media del *cluster* determinando infine quelli significativi con metodi di permutazione. Jenny Bryan *et al* rif [34] descrivono un metodo per selezionare i geni più significativi nel caso in cui il *dataset* sia composto da un gruppo di p geni e si abbiano due osservazioni per ogni soggetto oppure i soggetti appartengano a due popolazioni. Notano che la varianza campionaria e la presenza di geni poco correlati hanno un forte impatto sull'algoritmo di *clustering* e sulle misure di correlazione interna . Viene discusso un metodo per risolvere questo problema come l'applicazione di una procedura di analisi di raggruppamento che incorpora assieme passi di partizione e di aggregazione. Presentano anche un metodo per selezionare geni differenzialmente espressi

senza fare uso di assunti distributivi. Lazzeroni e Art Owen (2002) rif [35] presentano i *plaid models* che sono una forma di doppia analisi *cluster* che permette ad un gene di appartenere a più di un gruppo oppure a nessuno e permette anche ad un *cluster* di essere definito rispetto ad un solo sottogruppo di geni; la sovrapposizione dei *cluster plaid model* incorpora anche modelli di tipo ANOVA additivi all'interno di ciascun doppio *cluster*. Usando questi modelli sono riusciti a dare una struttura interpretabile ai geni. Sempre nel 2002 Gengxin chen *et al* rif [36] applicano diversi algoritmi di analisi di raggruppamento su un *dataset* di espressioni geniche di cellule embrionali. Propongono diversi indici basati sull'omogeneità interna, sulla separabilità, sulle *shilouette,* sui geni in eccedenza in un dato gruppo ecc. I risultati dimostrano che il *dataset* pone effettivamente dei problemi per l'analisi *cluster*, gli autori valutano vantaggi e svantaggi dei vari algoritmi rispetto da alcune misure di qualità interne ed esterne ai gruppi. Lo studio fornisce quindi una linea generale su come scegliere tra diversi algoritmi e può aiutare ed estrarre dal *dataset* le informazioni biologiche più significative. Sandrine Doudoit *et al* (2002) rif [37] confrontano la performance di diversi metodi discriminanti (*nearest neighbor*, analisi discriminante lineare, alberi di classificazione, approcci *machine learning* come *bagging* e *bootsing*) che vengono vagliati su un *dataset* di leucemie. Nello stesso anno Efron e Tibshirani rif [38] affrontano due approcci inferenziali per un problema a due campioni in cui si è interessati a evidenziare la differenza tra soggetti trattati e non. Il primo è un metodo empirico Bayesiano che richiede una modellazione della probabilità a priori, il secondo è un approccio frequentista basato sul *false discovery rate* proposto da Beniamini e Hochberg nel 1995. Gil autori arrivano a concludere che i due modelli sono strettamente legati e possono essere applicati assieme per produrre inferenza simultanea. Ruffino *et al* rif [39] applicano congiuntamente metodi di apprendimento automatico (*machine learning*), selezione di variabili e *bagging* su tre *dataset* dimostrando che l'applicazione congiunta di SVM (*support vector machine*) e *bagging* può ridurre significativamente l'effetto

degli errori oltre a dimostrarsi particolaremente efficace nell'analisi di dati di espressione genica per l'individuazione di classi funzionali di geni e discriminazione tra tessuti sani e tumorali e tra diversi tipi di tumore. Nell'aprile dello stesso anno, Chris Ding rif [40] affrontano il problema della *class discovery* usando la *spectral graph partitioning metodologies*, presentano poi *l'optimal leaf node ordering algorthm* applicati sia ai metodi di analisi di raggruppamento gerarchici che agglomerativi. Katherine S. Pollard *et al* rif [41] definiscono una nuova misura di omogeneità interna ai *cluster*: il MSS (*mean split shilouette*). Propongono di scegliere il numero di gruppi (*cluster*) che minimizza il MSS dimostrando la validità della procedura su una *dataset* simulato di *microarray data.* Nel febbraio 2003 Romualdi, Campanaro *et al* rif [42] comparano diverse tecniche di supervised *cluster*ing sulla base della capacità di classificare correttamente diversi tipi di cancro usando inizialmente l'approccio della simulazione per controllare la grande variabilità tra ed entro i pazienti. Mettono a confronto diverse tecniche di riduzione della dimensionalità che andranno poi ad aggiungersi all'analisi discriminante e verranno comparate sulla base della loro capacità di catturare l'informazione genetica principale. I risultati della simulazione sono poi stati vagliati applicando gli algoritmi a due *datasets* di espressioni geniche di pazienti malati di cancro, misurando il corrispondente tasso di errata classificazione. Questo porta all'individuazione di un insieme di geni principalmente coinvolti nella caratterizzazione del tumore. Nel marzo 2003 Sullivan Pepe *et al* rif [43] distinguono tra tessuto canceroso e sano da un *dataset* derivante dall'espressione genica di 23 soggetti sani e 30 con tumore ovarico discutendo la scelta di misure statistiche per caratterizzare i geni differenzialmente espressi ed un metodo per quantificare il livello di confidenza nella classificazione dei geni del *dataset* analizzato. Nel marzo 2003 Erich Hungan *et al* rif [44] analizzano un campione di 89 pazienti con tumore della mammella usando tecniche non lineari allo scopo di mettere in luce modelli d'interazioni di gruppi di geni che hanno valore predittivo per singoli pazienti relativamente

alla presenza di linfonodi con metastasi e ricaduta nella malattia. Trovano dei pattern in grado di fare previsioni con un accuratezza del 90%. Nell'aprile dello stesso anno Michael O'Neill e Li Song rif [45] studiano le reti artificiali applicate su dati da *microarray* da pazienti con linfoma di tipo DLCL e, per la prima volta, prevedono con accuratezza del 100% il tempo di sopravvivenza, restringendo il profilo genico a meno di tre dozzine di geni per ogni classificazione. Identificano le reti artificiali come strumento migliore sia per l'individuazione di gruppi di geni sia per evidenziare i geni più importanti che producono una corretta classificazione dei pazienti. Susmita Datta e Samnath (2003) *et al* rif [46] considerano sei algoritmi di *clustering* e valutano la loro *performance* sulla base di un *dataset* normale e due *datasets* simulati riguardanti osservazioni temporali. Formulano tre strategie di validazione individuando un metodo (DIANA) che funziona meglio degli altri. Sebastiani *et al* rif [47] riassumono nel loro articolo i fondamenti della *microarray tecnology* e descrivono i principali problemi che nascono dall'analisi di questo tipo di dati Keller (2003) rif [48] dimostra l'equivalenza di alcuni modelli lineari che in letteratura appaiono come modelli molto diversi ma che in realtà dimostra essere riformulazioni dello stesso modello generale che costituisce l'approccio base per l'analisi dei dati da *microarray*.

Capitolo 2
I dati

2.1 Il dataset

Il *dataset* è stato fornito dal Centro di ricerca Interdipartimentale per le Biotecnologie Innovative (CRIBI) dell'Università degli studi di Padova. Contiene i profili genetici di 22 soggetti distinti in quattro gruppi secondo il tipo di leucemia da cui risultano affetti; si hanno pertanto

- 10 soggetti affetti da leucemia linfoblastica acuta-B (ALL-B)
- 5 soggetti affetti da leucemia linfoblastica acuta-T (ALL-T)
- 3 soggetti affetti da leucemia linfoblastica acuta traslocata (ALL-Tr)
- 4 soggetti affetti da leucemia mieloide acuta (AML)

Per ciascun soggetto si dispone dell'espressione di 4992 geni. In aggiunta è stata fornita una prima selezione di 218 geni, scelti dai ricercatori mediante procedure di filtrazione ed identificati come geni di una certa rilevanza. Nel seguito ci si riferirà al *dataset* di 4992 geni con l'espressione *dataset completo,* a quello di 218 geni con il termine *dataset ridotto.*

Entrambi sono costituiti da misure già normalizzate, ossia per ciascun gene si dispone del logaritmo in base due del rapporto delle misure di fluorescenza per ciascun canale (o tinta). Inoltre sono già state eseguite le procedure di normalizzazione di cui si è parlato nel paragrafo 1.3.3.

Per evitare fraintendimenti, nel resto dell'elaborato, quando non specificato diversamente, si parlerà di *soggetto* o *unità* con riferimento ai pazienti, mentre si useranno i termini *variabile, espressione o gene* per riferirsi

ai valori assunti da ciascun gene nei 22 soggetti. In varie occasioni infatti, i 4992 geni saranno trattati come unità statistiche sulle quali si dispone di 22 osservazioni.

Si rileva infine la presenza di valori mancanti, dovuti non alla mancata espressione di un gene ma a problemi od errori che possono essere insorti nella fase sperimentale o di computazione ed elaborazione dell'immagine, nonché nel processo di traduzione delle tinte in numeri. Complessivamente si hanno il 6.58% di dati mancanti distribuiti in varie percentuali tra le unità; circa il 66.3% dei geni ha almeno un valore mancante, mentre il 4.46% ne ha più della metà. La Figura 2.1 fornisce una prima idea relativamente alla loro collocazione all'interno del *dataset*. Si può notare che tre sono i soggetti con la maggiore concentrazione di dati mancanti: le unità numero 9, 19 e 20. In appendice si riporta una tabella riassuntiva (Tabella A.2.1) relativa alla proporzione di valori mancanti per soggetto, per gene e per tipo di patologia.

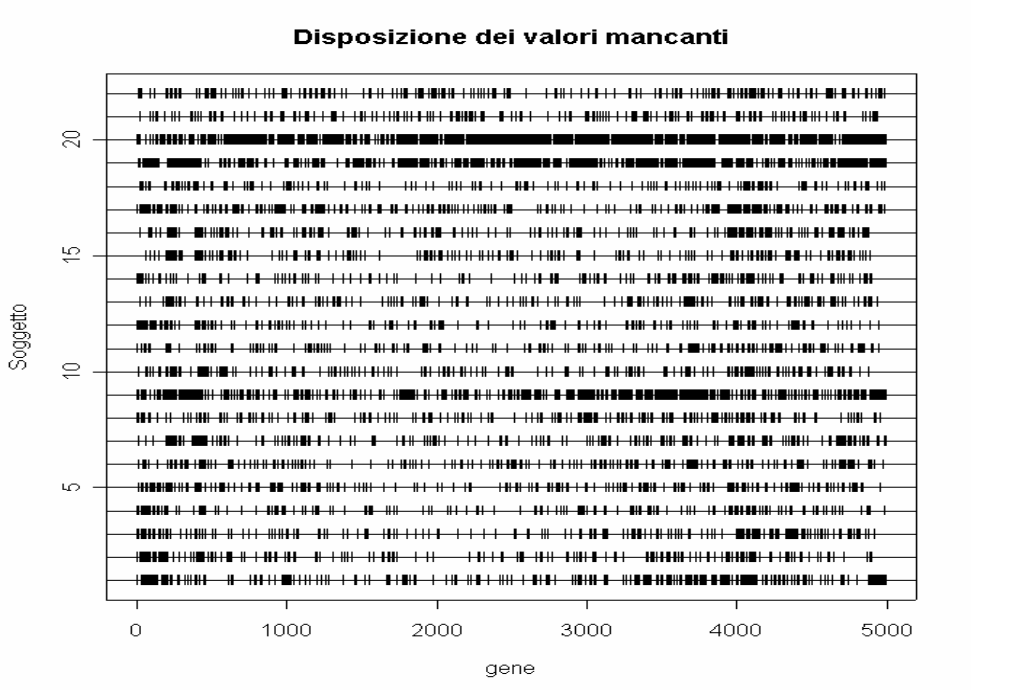

Figura 2.1 Disposizione dei valori mancanti all'interno della matrice dei dati: ciascuna riga rappresenta un paziente, i segmenti verticali indicano invece il gene per il quale non è stato rilevato il valore d'espressione. La figura è stata realizzata mediante la routine A2.3.r13.

2.2 Imputazione dei valori mancanti

Il primo problema che si incontra nell'analisi di dati derivanti da *microarray* è l'esistenza di valori mancanti nella matrice dei dati.

Le ragioni per cui spesso vengono a mancare le misure relative ad un certo gene sono diverse e possono essere connesse agli strumenti utilizzati (presenza di polvere o graffi nei vetrini…), oppure al trattamento computazionale dell'immagine e al processo di trasformazione del segnale luminoso in dato numerico (insufficiente risoluzione, alterazione dell'immagine…).

Sfortunatamente, la maggior parte degli algoritmi per l'analisi dell'espressione genica non prevede la presenza di dati mancanti, e diversi metodi come l'analisi *cluster* di tipo gerarchico o il metodo delle *k*-medie non sono robusti alla presenza degli stessi oppure perdono in efficienza già con una bassa percentuale (Tibshirani 2001). Sorge quindi il problema dell'imputazione dei dati mancanti.

A volte, il valore mancante del logaritmo in base due del rapporto viene sostituito con uno zero come se il gene non presentasse differenze di espressione tra il soggetto malato e quello sano. Un'operazione di questo genere può però essere fuorviante, in quanto si potrebbe additare un gene come differenzialmente espresso tra una patologia ed un'altra solo perché in un gruppo non si sono avute osservazioni per quel gene. Certamente dunque, molte tecniche di analisi trarrebbero beneficio da una stima più accurata dei valori mancanti.

2.1.1 Tecniche di imputazione

Di seguito verranno studiati quattro metodi di imputazione. In un momento successivo saranno messi a confronto per verificare quale di essi sia il più adeguato per il dataset che si sta analizzando. Le tecniche di imputazione che verranno analizzate sono quattro e prendono il nome di: "metodo della media

generale", "metodo della media di gruppo", "metodo della decomposizione in valori singolari" e "metodo *k-nearest neighbors*". Tali metodi prevedono la lettura del *dataset* come una matrice 4992×22 (i geni come unità statistiche), ossia un valore sarà considerato mancante non per il soggetto ma per il gene. È anche in ragione di questo motivo che si è ritenuto opportuno non procedere all'eliminazione dei tre soggetti che presentano un'elevata percentuale di dati mancanti. Nel seguito saranno illustrati nel dettaglio i quattro metodi e verranno presentati i risultati salienti relativi alla loro applicazione.

- **Metodo d'imputazione con la media generale.** Il metodo prevede la stima del valore mancante con la media dell'espressione assunta dallo stesso gene su tutti i soggetti (la media generale per l'appunto). Questa tecnica, così come quella di sostituzione del valore zero, non è ottimale in quanto non tiene in considerazione la struttura di correlazione dei dati, ossia la stima di un valore mancante in certo gene non usa eventuali informazioni che potrebbero provenire da geni ad esso correlati e che consentirebbero una stima più accurata. La procedura d'imputazione è implementata nella *routine* A2.3.r1.

- **Metodo d'imputazione con la media di gruppo.** Il metodo stima il valore mancante di un gene per un paziente con la media dello stesso gene osservata sui soli pazienti che appartengono al medesimo gruppo (ossia affetti dalla stessa forma di leucemia). Un problema nel quale ci si può imbattere con questa tecnica è la stima dell'espressione di quei geni per i quali non si è rilevata alcuna misura su un intero gruppo. In tal caso si può procedere sostituendo il valore mancante con la media delle espressioni dello stesso gene sui restanti gruppi. La procedura d'imputazione è stata realizzata con le *routines* A2.3.r2.

- **Metodo d'imputazione con la decomposizione in valori singolari (o metodo SVD).** Tale metodo (Tibshirani *et al*. rif [51]) si serve della decomposizione in valori singolari (SVD) e crea un insieme di vettori

mutuamente ortogonali che vengono poi combinati linearmente per approssimare le espressioni di tutti i geni del *dataset* X. Sia dunque

$$X_{p\times n} = U_{p\times p} \quad D_{p\times n} \quad V^T_{n\times n}$$

la decomposizione in valori singolari di X, dove D è una matrice *pseudodiagonale* che riporta sulla diagonale principale gli autovalori di X, mentre U e V^T sono matrici ortogonali contenenti i corrispondenti autovettori. Si consideri ora la decomposizione troncata

$${}_J\hat{X} = {}_JU_{p\times p} \quad {}_JD_{p\times J} \quad {}_JV^T_{J\times n}$$

dove ${}_JD$ è la matrice *pseudodiagonale* contenente i primi $J = n$ autovalori di X, ${}_JU$ e ${}_JV^T$ le corrispondenti matrici di autovettori. Questa combinazione fornisce la matrice $\hat{X}_J$ di rango J che meglio approssima X tra tutte le matrici di rango J, ossia risolve il problema di minimo vincolato

$$\min_{\mathrm{rk}(M)=J} \| X - M \|^2 .$$

dove M è una matrice di rango J.

La soluzione può essere interpretata anche dal punto di vista della regressione ai minimi quadrati ordinari (MQO). Sia infatti x un qualsiasi vettore riga di X e si consideri la regressione MQO degli n elementi di x sui J autovettori di ${}_JV$, ciascuno di lunghezza n. Regredire x su ${}_JV$ equivale a risolvere il problema di minimo

$$\min_{ß}(x - {}_JVß)^2$$

che ha come soluzione $\hat{ß} = \left({}_JV^T\,{}_JV\right)^{-1}{}_JV^Tx$ che equivale a $\hat{ß} = {}_JV^Tx$ essendo ${}_JV$ ortogonale. I valori previsti saranno pertanto $\hat{x} = {}_JV\hat{ß}$. Generalizzando per tutte le righe di X si ha che $X\,{}_JV = {}_JU\,{}_JD$ fornisce tutti i coefficienti per ogni riga di X mentre $\hat{X} = {}_JU\,{}_JD\,{}_JV^T$ dà tutti i valori stimati. Così, una volta individuati i J autovalori più significativi e definita ${}_JV^T$, la decomposizione in

valori singolari approssima ciascuna riga di X. Si noti che la decomposizione in valori singolari può essere condotta solo su una matrice priva di valori mancanti. Pertanto si è partizionato il *dataset* X in due matrici X^c ed X^m; la prima (1681×22) contiene tutti e soli i geni che non presentano valori mancanti, la seconda (3311×22) raccoglie invece i geni che possiedono almeno un valore mancante. Indicando con x^* un qualsiasi vettore appartenente alla matrice X^m, i suoi valori mancanti possono essere imputati con un procedimento di regressione simile che sfrutta la decomposizione in valori singolari di X^c. Una volta identificati i J autovalori più significativi, si stima il dato mancante k nel gene i, prima regredendo questo gene sui J autovettori, e poi usando il coefficiente di regressione per ricostruire k da una combinazione lineare dei J autovettori. Naturalmente questa operazione prevede che il k-esimo valore del gene i e gli elementi k-esimi dei J autovettori, non vengano usati nei calcoli per la stima del coefficiente di regressione che servirà per la stima del k-esimo valore mancante. In particolare si cercherà il vettore di parametri b che risolve il seguente problema di minimo vincolato

$$\min_{ß} \sum_{l\ nonmancanti} \left(x_i^* - \sum_{j=1}^{J} v_{ij}\, ß_j \right).$$

Dopo aver applicato alla matrice X^c la decomposizione in valori singolari, si definisce ${}_JV^{*T}$ la matrice ottenuta rimuovendo da ${}_JV^T$ le colonne che corrispondono agli elementi mancanti di x^* e ${}_JV^{(*)T}$ la matrice ad essa complementare. La soluzione al problema di minimo sarà pertanto $\hat{ß} = {}_JV^{*T} x^*$, ed i valori mancanti potranno essere stimati con ${}_JV^{(*)T}\hat{ß}$. Questa procedura d'imputazione è stata implementata nelle *routines* A2.3.r3-A2.3.r4-A2.3.r5-A2.3.r7.

- **Metodo d'imputazione *k-nearest neighbors* (metodo o *Knn*).** Tale metodo prevede l'utilizzo, nella stima dei dati mancanti, dei geni che più somigliano a quello che presenta il valore mancante. Ideata da Tibshirani *et al.*

(rif [51]), questa tecnica prevede ancora la partizione del *dataset* nelle due matrici X^c ed X^m. Indicato con x^* un qualsiasi vettore appartenente alla matrice X^m, l'algoritmo consiste nel calcolo della distanza tra x^* e tutti i geni contenuti nella matrice X^c usando solo le coordinate dei valori di x^* che non risultano mancanti. Successivamente il dato da imputare viene calcolato con una media pesata dei k geni che risultano più prossimi (o "vicini") ad x^*. In questa media, il contributo di ciascun gene viene pesato in base alla somiglianza della sua espressione con quella del gene x^*. In particolare si sono usati pesi inversamente proporzionali alla distanza e tali da sommare a uno. La metrica che viene suggerita per il calcolo delle distanze è la quella Euclidea che risulta sufficientemente accurata nonostante sia spesso sensibile agli *outliers* che possono presentarsi nei dati derivanti da *microarray*. Infatti sembra che, applicando la trasformata logaritmica, venga notevolmente ridotto l'effetto dei valori anomali nella determinazione di misure di similarità. La procedura d'imputazione è stata realizzata con la *routine* A2.3.r4-A2.3.r5-A2.3.r6 e A2.3.r7.

Per quanto concerne la scelta dei parametri J e k che regolano il numero di autovalori nel caso dell'imputazione SVD ed il numero di geni "vicini" da considerare nel calcolo della media pesata nel caso *Knn*, si è cercato di generare un pattern di (finti) valori mancanti nella matrice X^c che potesse somigliare a quello della matrice iniziale. A tal fine si sono campionate 1681 righe da X e si sono assegnati dei valori mancanti nelle 1681 righe di X^c usando la stessa collocazione che avevano nelle righe selezionate. Per questa nuova versione di X^c si sono applicati gli algoritmi di imputazione per un insieme di valori di J e di k e, per ciascun valore all'interno di questo insieme, si è calcolato un indice di bontà dell'imputazione (RMS error, Root Mean Squared error) definito come la media dei quadrati degli scarti tra valore vero e valore stimato, il tutto sotto radice:

$$RMS\ error = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(x_i - \hat{x}_i\right)^2}$$

dove *m* è il numero di valori mancanti che sono stati assegnati alla matrice X^c, mentre x_i ed $\hat{x}_i$ sono rispettivamente il vero valore ed il valore stimato del finto dato mancante. Questa operazione è stata effettuata su cinque diversi campioni allo scopo di scegliere il valori di *k* e di J per i quali l'RMS error risultasse minimo. Le Figura 2.2 eFigura 2.3 e riportano l'RMS error in funzione di *J* e di *k* rispettivamente; ogni linea rappresenta ciascuno dei cinque campioni per i quali si è applicato l'algoritmo (queste analisi sono state realizzare con le *routines* A2.3.r8-A2.3.r9)

Per quanto riguarda la scelta di *J*, il grafico della Figura 2.2 mette in luce che l'RMS error presenta un punto di minimo per *J* pari a 17-18, si è deciso di adottare *J*=17 che corrisponde alla scelta di usare all'incirca il 85% degli autovettori.

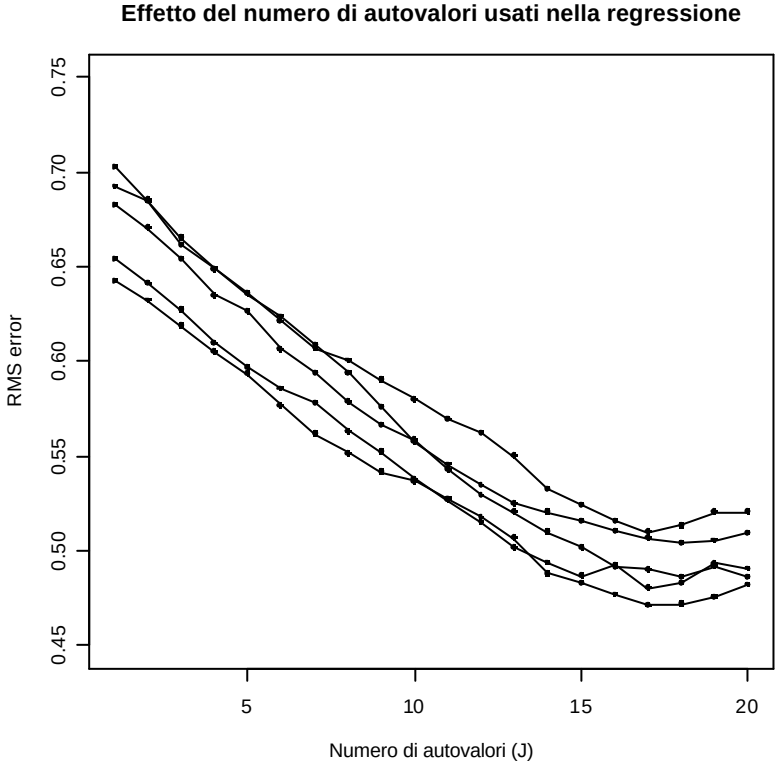

Figura 2.2 Metodo SVD: effetto del parametro *J* che regola il numero di autovettori usati per la stima del coefficiente di regressione ß sull'RMS error. Ciascuna curva corrisponde ad uno dei campioni su cui si è applicato l'algoritmo.

L'RMS error in funzione di *k*, invece, ha un minimo piuttosto piatto in corrispondenza dell'insieme di valori che vanno da 5 a 13-14. Pertanto è parso sensato adottare un valore di *k* pari a 10.

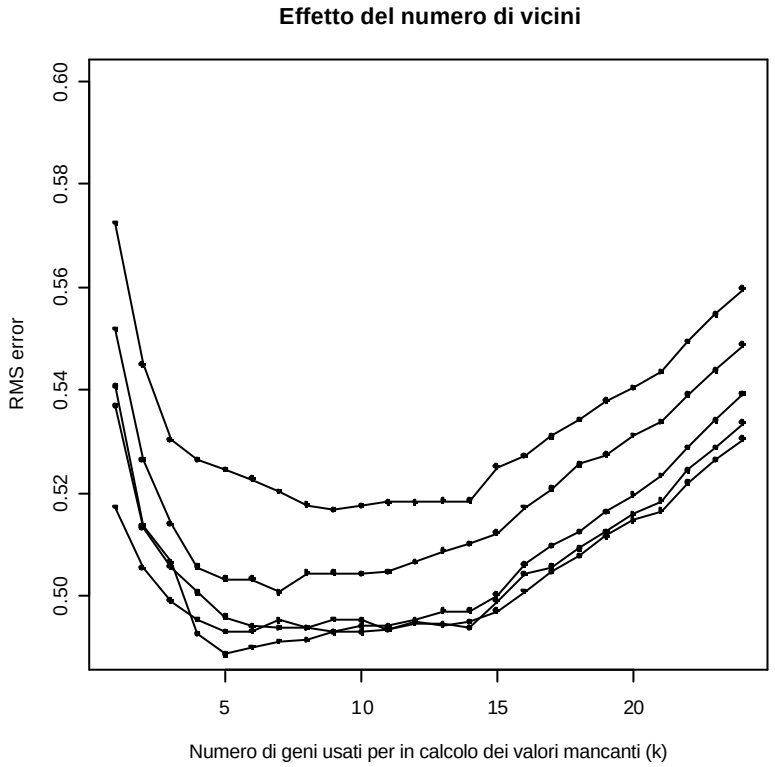

Figura 2.3 Metodo *Knn*: andamento dell'RMS error in funzione del coefficiente *k* che regola il numero di geni su cui si calcola la media pesata per l'imputazione dei dati mancanti. Le varie curve corrispondono a diversi *dataset* su cui è stato applicato l'algoritmo.

Il grafico in Figura 2.3 mette in evidenza come la prestazione dell'algoritmo declini quando si usano pochi valori per la stima, e ciò è dovuto alla eccessiva enfasi di alcuni modelli di espressione. Il peggioramento della *performance* per valori crescenti di *k* può essere spiegato in ragione a due motivi. Innanzitutto

l'inclusione di modelli d'espressione significativamente diversi (o lontani) dal gene d'interesse può far perdere in accuratezza in quanto il gruppo di geni su cui si calcola la media pesata è troppo grande e non rilevante per la stima; secondariamente è importante tener presente che nei dati derivanti da *microarray* può esserci molto rumore, e con l'aumentare di k il contributo del disturbo alla stima supera il contributo del segnale, causando una diminuzione dell'accuratezza.

Si noti infine che con l'aumentare di k ci si avvicina al primo metodo di stima (quello che prevede la sostituzione della media globale). Si può pertanto intuire che l'accuratezza di questo metodo è superiore a quella del metodo sopraccitato.

Nella Figura 2.4 si riportano infine gli errori RMS relativi ai quattro metodi di imputazione ciascuno dei quali è stato applicato a cinque diversi campioni. Come anticipato, per i metodi *Knn* ed SVD i parametri k e J si sono fissati rispettivamente a 10 e 17. È evidente la superiore accuratezza dei metodi *Knn* ed SVD mentre stupisce la scarsa *performance* del metodo delle medie di gruppo rispetto al metodo della media generale. Nel resto dell'analisi si adotterà il metodo *Knn* in quanto è il più indicato per il *dataset* di cui si dispone, mentre in metodo SVD viene consigliato per serie storiche di espressioni geniche, ossia quando si è interessati a studiare l'evoluzione dell'espressione di un gene nel tempo al variare di alcune condizioni.

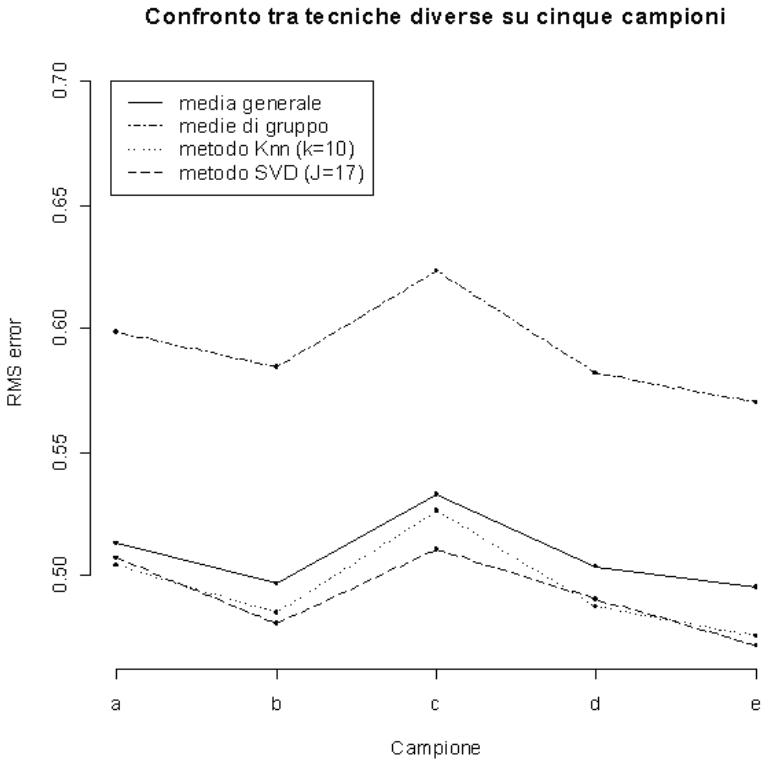

Figura 2.4 Confronto tra le performance dei quattro metodi d'imputazione che sono stati applicati sugli stessi cinque campioni, Ciascuna curva rappresenta un singolo metodo. Per gli algoritmi *Knn* ed SVD, i parametri *k* e *J* sono stati fissati rispettivamente a 10 e 17

2.1.2 Efficacia delle tecniche di imputazione

Nel paragrafo precedente, si è verificata l'efficacia di quattro metodi di imputazione dei dati mancanti mediante l'utilizzo dell'errore RMS. Nel seguito, l'efficacia dei metodi verrà valutata con riferimento al processo di classificazione dei pazienti in gruppi omogenei. L'obiettivo è di verificare se i quattro metodi d'imputazione possono avere un impatto diverso sulla formazione di sottogruppi omogenei di pazienti. L'analisi è stata limitata ad algoritmi di classificazione basati su partizioni delle osservazioni.

Si sono quindi applicati (previa standardizzazione delle variabili) il metodo delle *k-medie* e dei *medoidi* per creare quattro gruppi di pazienti che fossero omogenei sulla base dell'espressione dei loro geni. È stata poi verificata la rispondenza con la reale suddivisione dei soggetti a seconda delle quattro forme di leucemia e la stabilità della classificazione rispetto a diversi metodi di imputazione. Il metodo delle *k-medie*, ideato da MacQueen (1967), consta nella suddivisione del campione in un numero *k* di gruppi prefissato. L'algoritmo, iterativo, prevede il calcolo del vettore delle medie di gruppo (o centroide) e dell'allocazione di ogni unità al gruppo con il centroide più vicino. L'iterazione continua fintantoché l'algoritmo si stabilizza e nessuna unità viene assegnata ad un gruppo diverso da quello a cui apparteneva nel passo precedente. Il metodo dei *medoidi* (si veda L. Kaufman and P. J. Rousseeuw, (1990) rif [50]) è analogo a quello delle *k*-medie, solo che, ad ogni passo, prevede anziché il calcolo della media di gruppo, l'identificazione del *medoide* ossia dell'osservazione più rappresentativa del gruppo; successivamente ogni osservazione viene allocata al gruppo con il medoide più vicino e l'algoritmo si ripete finché non giunge a convergenza.

Queste tecniche sono state applicate ai quattro *dataset* derivanti dai quattro diversi metodi di imputazione, per valutare come e se questi ultimi influiscano sul processo di classificazione e sui risultati delle analisi. Per entrambi gli algoritmi di classificazione si è scelto di formare quattro gruppi idealmente corrispondenti alle quattro forme di leucemia.

Si noti che, la funzione `kmeans` della libreria `mva`, che si è utilizzata per l'applicazione del metodo della *k*-medie, richiede in *input* il numero *k* di gruppi in cui devono essere suddivise le unità o, in alternativa, *k* vettori che vengono utilizzati come centro di ciascun *cluster* quando l'algoritmo viene inizializzato. Nel caso in cui si fornisca soltanto il numero di gruppi in cui devono essere suddivise le unità, la funzione sceglie a caso *k* righe del *dataset* che vengono utilizzate come centro iniziale dei *k* gruppi. Adottando questa soluzione, però,

l'algoritmo non converge sempre alla stessa soluzione e presenta risultati diversi a seconda dell'insieme iniziale di righe che seleziona. Questo probabilmente accade a causa della presenza di minimi locali. Si è pertanto reso necessario passare all'algoritmo quattro vettori da utilizzare come punto di partenza; in particolare si sono fornite le medie di gruppo. Fatta questa premessa, è chiaro che non avrebbe senso impostare un confronto fra i risultati del metodo delle *k*-medie e dei medoidi, in quanto, nel caso in esame, il primo ha a disposizione più informazioni rispetto al secondo.

La Tabella 2.1 riassume risultati ottenuti sul *dataset* completo: le unità contrassegnate dallo stesso numero sono state attribuite al medesimo gruppo; non esiste invece relazione tra unità con lo stesso numero ma su righe diverse.

Per entrambi i metodi, gli esiti della classificazione non lusingano, in quanto i gruppi generati dall'algoritmo non riproducono i quattro gruppi definiti dalle quattro forme di patologia di cui sono affetti i pazienti. Per quanto riguarda l'impatto delle tecniche di imputazione sugli esiti della classificazione, si nota che per entrambi gli algoritmi i risultati non variano a seconda della tecnica utilizzata, ossia le unità vengono raggruppate allo stesso modo qualunque sia il metodo di imputazione applicato.

Metodi di clustering ?		Metodo delle k-medie				Metodo dei medoidi			
Metodi di imputazione ?		media generale	medie di gruppo	SVD	Knn	media generale	medie di gruppo	SVD	Knn
Reale gruppo di appartenenza del soggetto (patologia)	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	2	2	2	2	2	2	2	2
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	3	3	3	3	3	3	3	3
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	3	3	3	3	3	3	3	3
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	3	3	3	3	2	2	2	2
	ALL-T	1	1	1	1	2	2	2	2
	ALL-T	3	3	3	3	3	3	3	3
	ALL-Tr	3	3	3	3	3	3	3	3
	ALL-Tr	3	3	3	3	3	3	3	3
	ALL-Tr	3	3	3	3	3	3	3	3
	AML	3	3	3	3	2	2	2	2
	AML	3	3	3	3	3	3	3	3
	AML	4	4	4	4	4	4	4	4
	AML	3	3	3	3	3	3	3	3

Tabella 2.1 Confronto tra i quattro metodi di imputazione analizzati: risultati relativi all'applicazione dei metodi delle *k*-medie e dei medoidi per la classificazione dei soggetti su *datasets* derivanti dall'applicazione dei diversi algoritmi di imputazione (*dataset* completo). Le unità contraddistinte dallo stesso numero sono state assegnate allo stesso raggruppamento; le unità con lo stesso numero ma su righe diverse non stanno tra loro in alcuna relazione

Le stesse analisi, condotte sui 218 geni selezionati[2], forniscono risultati nettamente più confortanti (Tabella 2.2), in quanto buona parte delle osservazioni viene ora classificata correttamente. Con il metodo di imputazione delle medie di gruppo e con quello della media generale associato al metodo dei medoidi addirittura non si commettono errori. I risultati si discostano da quelli derivanti dall'analisi con tutti i geni anche perché in questo caso le

[2] Per la stima dei valori mancanti nel *dataset* ridotto si è preferito riapplicare i quattro algoritmi (su questo dataset) anziché attuare una selezione dei geni dal dataset completo sul quale erano già state applicate le imputazioni.

aggregazioni variano leggermente a seconda del metodo di imputazione. Inoltre, eccezion fatta per il metodo della media generale, i raggruppamenti effettuati dal metodo dei medoidi sono identici a quelli generati dal metodo delle *k*-medie.

Metodi di clustering ?		Metodo delle k-medie				Metodo dei medoidi			
Metodi di imputazione ?		media generale	medie di gruppo	SVD	Knn	media generale	medie di gruppo	SVD	Knn
Reale gruppo di appartenenza del soggetto (patologia)	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	3	1	3	3	1	1	3	3
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	2	2	2	2	2	2	2	2
	ALL-Tr	3	3	3	3	3	3	3	3
	ALL-Tr	3	3	3	3	3	3	3	3
	ALL-Tr	3	3	3	3	3	3	3	3
	AML	4	4	4	4	4	4	4	4
	AML	4	4	4	4	4	4	4	4
	AML	4	4	4	4	4	4	4	4
	AML	4	4	4	4	4	4	4	4

Tabella 2.2 Confronto tra i quattro metodi di imputazione analizzati: risultati relativi all'applicazione dei metodi delle *k*-medie e dei medoidi per la classificazione dei soggetti su *datasets* derivanti dai diversi algoritmi di imputazione (*dataset* completo). Le unità contraddistinte dallo stesso numero sono state assegnate al medesimo raggruppamento.

Si noti inoltre che l'ultimo soggetto del gruppo ALL-B è l'unico a risentire spesso di un'errata classificazione (sottolineato in tabella) ed in tutte queste occasioni, non solo viene allocato assieme alle unità che appartengono al gruppo ALL-Tr, ma è anche il medoide rappresentante di questo gruppo. È pertanto probabile che si tratti di un'osservazione anomala o comunque di un'unità da tenere sotto controllo nelle successive analisi.

Lo stesso studio è stato condotto su variabili non sottoposte a standardizzazione; le tabelle corrispondenti vengono riportate in appendice (Tabella A.2.2 e Tabella A.2.3). I risultati suggeriscono grossomodo le stesse considerazioni anche se, ponendoli a confronto con quelli presentati sopra, si intuisce che per questo tipo di analisi è consigliabile standardizzare le variabili in quanto ciò consente di ottenere risultati meno sensibili al metodo di imputazione ed un raggruppamento più simile alla reale suddivisione dei soggetti nei quattro gruppi di patologia.

Da questa analisi emergono due conclusioni principali. Dal punto di vista dell'impatto di diverse tecniche di imputazione sulle analisi che possono essere condotte, si è verificato che queste tecniche non sembrano essere molto sensibili rispetto a diversi metodi di imputazione. Dal punto di vista applicativo, si è verificata una certa difficoltà nel riprodurre la reale suddivisione dei pazienti nei quattro sottogruppi di patologia. Questa considerazione vale soprattutto per il *dataset* completo in cui sembra esistere molto rumore che rende difficile l'identificazione dei quattro gruppi di patologia. Per quanto concerne il *dataset* ridotto, si è visto che l'ultima unità del gruppo ALL-B risente speso di un'errata classificazione in quanto viene allocata assieme alle unità del gruppo ALL-Tr. Questo tipo di anomalia è stata riscontrata anche da parte dei ricercatori dal CRIBI che hanno sottoposto questi dati alle prime analisi. Anche su consiglio di questi ultimi si è ritenuto opportuno non eliminare dal *dataset* il paziente in questione. Essi, infatti, hanno escluso che tale anomalia potesse derivare solo da errori nella fase sperimentale in quanto le misure sui livelli di espressione sono state ottenute da una media su quattro osservazioni ripetute e, se si fossero verificati errori tali da compromettere il risultato finale, il paziente sarebbe stato eliminato dal *dataset* da parte dei ricercatori stessi.

Appendice

A2.1 Valori mancanti

Tabella riassuntiva dei valori mancanti, per gene soggetto e patologia. La percentuale di valori mancanti per soggetto è attorno al 4,5%, fanno eccezione quattro unità, tre delle quali (n° 9, 19 e 20) presentano percentuali piuttosto elevate. Si noti che due di queste appartengono al gruppo AML, patologia per la quale si ha la più elevata percentuale di dati mancanti.

Valori mancanti per soggetto			
Patologia	Unità	Numero	Percentuale
ALL-B	1	328	6,571
ALL-B	2	216	4,327
ALL-B	3	207	4,147
ALL-B	4	208	4,167
ALL-B	5	238	4,768
ALL-B	6	225	4,507
ALL-B	7	233	4,667
ALL-B	8	210	4,207
ALL-B	9	677	13,562
ALL-B	10	233	4,667
ALL-T	11	219	4,387
ALL-T	12	201	4,026
ALL-T	13	223	4,467
ALL-T	14	203	4,067
ALL-T	15	223	4,467
ALL-Tr	16	234	4,688
ALL-Tr	17	328	6,571
ALL-Tr	18	196	3,926
AML	19	951	19,050
AML	20	1270	25,441
AML	21	216	4,327
AML	22	220	4,407

Percentuale di valori mancanti per patologia	
ALL-B	5,559
ALL-T	4,283
ALL-Tr	5,061
AML	13,306

Percentuale e numero di geni con x valori mancanti		
x	Percentuale	numero
1	33,674	1681
2	29,728	1484
3	17,107	854
4	9,936	496
5	4,788	239
6	1,923	96
7	1,082	54
8	0,481	24
9	0,280	14
10	0,321	16
11	0,240	12
12	0,100	5
13	0,120	6
14	0,160	8
15	0,020	1
16	0,040	2

Tabella A.2.1 Tabella riassuntiva dei valori mancanti. La tabella è stata realizzata con le *routines* A2.3.r10-A2.3.r11-A2.3.r12.

A2.2 Efficacia delle tecniche d'imputazione

Applicazione del metodo delle *k*-medie e dei medoidi sui quattro *dataset* che corrispondono ai quattro metodi di imputazione condotti su variabili non standardizzate. La prima tabella (Tabella A.2.2) si riferisce al *dataset* completo, la seconda (Tabella A.2.3) a quello ridotto. Anche in questo caso, come riscontrato nel corso del capitolo, entrambi gli algoritmi sono abbastanza robusti rispetto al metodo di imputazione (in quanto i raggruppamenti creati da ciascuno dei due metodi non variano molto a seconda della tecnica di imputazione). Infatti, sono soltanto tre le unità (sottolineate in tabella) che cambiano il gruppo di appartenenza a seconda del metodo di imputazione. Si noti infine che la penultima unità del gruppo AML viene sempre isolata in un gruppo a se stante e che, come succede nel caso di variabili standardizzate, la decima unità del gruppo ALL-B viene allocata nel gruppo che contiene le unità del gruppo ALL-Tr.

Metodi di clustering ?		Metodo delle k-medie				Metodo dei medoidi			
Metodi di imputazione ?		media generale	medie di gruppo	SVD	Knn	media generale	medie di gruppo	SVD	Knn
Reale gruppo di appartenenza del soggetto (patologia)	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	2	2	2	2	2	2	2	2
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	3	3	3	3	3	3	3	2
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	1	1	1	1	2	2	2	2
	ALL-B	3	3	3	3	3	3	3	3
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	2	2	2	2	2	2	2	2
	ALL-T	2	2	3	2	2	2	2	2
	ALL-Tr	3	3	3	3	3	3	3	3
	ALL-Tr	3	3	3	3	3	3	3	3
	ALL-Tr	3	3	3	3	2	2	3	2
	AML	3	3	3	3	2	2	2	2
	AML	3	3	3	3	3	3	3	3
	AML	4	4	4	4	4	4	4	4
	AML	3	3	3	3	3	3	3	3

Tabella A.2.2 Confronto tra metodi di imputazione: risultati relativi all'applicazione dei metodi delle *k*-medie e dei medoidi per la classificazione dei soggetti su *datasets* derivanti dai diversi algoritmi di imputazione (*dataset* completo). Le unità contraddistinte dallo stesso numero sono state assegnate allo stesso raggruppamento; le unità con lo stesso numero ma su righe diverse non stanno tra di loro in alcuna relazione

Per quanto concerne il *dataset* ridotto, buona parte delle osservazioni viene ora classificata correttamente e con il metodo di imputazione delle medie di gruppo non si commettono errori. Anche in questo caso, i risultati si differenziano da quelli derivanti dall'analisi con tutti i geni anche perché ora i raggruppamenti variano a seconda del metodo di imputazione ma, fissato quest'ultimo, i raggruppamenti non variano tra metodo dei medoidi e delle *k*-medie. Si noti infine che, anche in questo caso, l'ultimo soggetto del gruppo ALL-B risente spesso di un'errata classificazione, mentre, in due occasioni, viene a costituirsi un gruppo che contiene solo la penultima unità della forma AML.

Metodi di clustering ?		Metodo delle k-medie				Metodo dei medoidi			
Metodi di imputazione ?		media generale	medie di gruppo	SVD	Knn	media generale	medie di gruppo	SVD	Knn
Reale gruppo di appartenenza del soggetto (patologia)	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	1	1	1	1	1	1	1	1
	ALL-B	2	1	2	2	2	1	2	2
	ALL-T	3	2	3	3	3	2	3	3
	ALL-T	3	2	3	3	3	2	3	3
	ALL-T	3	2	3	3	3	2	3	3
	ALL-T	3	2	3	3	3	2	3	3
	ALL-T	3	2	3	3	3	2	3	3
	ALL-Tr	2	3	2	2	2	3	2	2
	ALL-Tr	2	3	2	2	2	3	2	2
	ALL-Tr	2	3	2	2	2	3	2	2
	AML	2	4	4	4	2	4	4	4
	AML	2	4	4	4	2	4	4	4
	AML	4	4	4	4	4	4	4	4
	AML	2	4	4	4	2	4	4	4

Tabella A.2.3 Confronto tra metodi di imputazione: risultati relativi all'applicazione dei metodi delle k-medie e dei medoidi per la classificazione dei soggetti su datasets derivanti dai diversi algoritmi di imputazione (dataset completo). Le unità contraddistinte dallo stesso numero sono state assegnate al medesimo raggruppamento.

A2.3 *Routines* R

A2.3.r1 `#sostituzione dei valori mancanti con la media generale`

```
sost.mean<-function (data)
{
dati<-data
m<-apply(data,1,mean,na.rm=T)
for (j in 1:ncol(data)){
    for (i in 1:nrow(data)){
        if (is.na(data[i,j]))
           dati[i,j]<-m[i]
    }
}
dati
}
```

A2.3.r2 `#sostituzione dei valori mancanti con la media di gruppo`

```
sost.mg<-function(dat){
dat<-as.matrix(dat)
Ana<-dat[,1:10]
Bna<-dat[,11:15]
Cna<-dat[,16:18]
Dna<-dat[,19:22]
Ana<-as.matrix(Ana)
Bna<-as.matrix(Bna)
Cna<-as.matrix(Cna)
Dna<-as.matrix(Dna)

#creazione dei vettori medie di gruppo

medA<-apply(Ana,1,mean,na.rm=T)
mA<-matrix(medA,nrow=1)
medB<-apply(Bna,1,mean,na.rm=T)
mB<-matrix(medB,nrow=1)
medC<-apply(Cna,1,mean,na.rm=T)
mC<-matrix(medC,nrow=1)
medD<-apply(Dna,1,mean,na.rm=T)
mD<-matrix(medD,nrow=1)

#sostituzione dei dati mancanti

a.mg<-mancanti(Ana,mA)
b.mg<-mancanti(Bna,mB)
c.mg<-mancanti(Cna,mC)
d.mg<-mancanti(Dna,mD)
dati.mg<-cbind(a.mg,b.mg,c.mg,d.mg)

dati.mg<-sost.mean(dati.mg)
dati.mg
```

```
}
```

A2.3.r3

```
svd.imput<-function (dat,xm,xmc,mat,j)          #dat
 sarebbe xc
{
mat<-matrix(mat,ncol=1)
data<-as.matrix(dat)
y<-1
x<-0
el<-NULL
for (i in 1:j){
    x<-x+1
    el[i]<-ncol(dat)-j+x
    }
dat<-as.matrix(dat)
dati<-as.matrix(dat)

dati.svd<-svd(dati)

VT<-t(as.matrix(dati.svd$v))
V<-as.matrix(dati.svd$v)
for (h in 1:nrow(xm)){
    missing<-NULL
    m<-0
    ind<-NULL
    l<-0
    for (k in 1:ncol(xm)){
         m<-m+1
         if (is.na(xm[h,k])){
             l<-l+1
             ind[l]<-m
            }
         }
         x.min<-matrix(xm[h,-ind],ncol=1)
         VT.min<-VT[-el,-ind]
         VT.compl<-VT[-el,ind]
         beta<- (VT.min %*% x.min)
         missing<-t(VT.compl) %*% beta

         f<-0
         for (z in 1:ncol(xm)){

         if (is.na(xm[h,z])){
             f<-f+1
             xm[h,z]<-missing[f]
             }
         }
}
da.def<-riordina(xm,dat,mat)
da.def

}
```

A2.3.r4

```
#crea la matrice Xc
```

```
togli<-function(data,mat){
x<-rep(0,22)
for (i in 1:nrow(mat)){
    if (mat[i,1]==0){
       x<-rbind(x,data[i,])
       }
}
x<-x[-1,]
x
}
```

A2.3.r5 `#crea la matrice Xm`

```
restante<-function(data,mat)
{
x<-rep(0,22)
for (i in 1:nrow(mat)){
if (mat[i,1]!=0){
x<-rbind(x,data[i,])
}
}
x<-x[-1,]
x
}
```

A2.3.r6 `#imputazione` *Knn*

```
elim.na<-function(xm,xc,k,mat)
{
xm2<-xm
library(fields)
for(t in 1:nrow(xm)){
   h<-0
   j<-0
   y<-NULL
   x<-0
   for (i in 1:ncol(xm)){      #Elimina sia da Xc che da
Xm le colonne
                               #corrispondenti ai NA
       x<-x+1
       if(is.na(xm[t,i])){
         j<-j+1
         y[j]<-x
         }
       }
   g<-(-y)
   xmm<-xm[t,g]
   xcc<-xc[,g]
   xmm<-as.matrix(xmm)
   xcc<-as.matrix(xcc)
   xmm<-t(xmm)
```

```
   #Crea un vettore di pesi inversamente proporzionale
   alla distanza
   dis<-NULL
   d<-NULL
   #calcola la distanza tra xmm e ciascuna riga di xcc
   for (r in 1:nrow(xcc)){
       dis<-(xmm-xcc[r,])**2
           d[r]<-sum(dis)
           d[r]<-sqrt(d[r])
       }

    d<-sort(d,index.return=T)   #ordinanento crescente
    delle distanze
                                # mantenendo gli indici
   dx<-as.matrix(d$x)
   dix<-as.matrix(d$ix)
   d1<-dx[1:k]
   d1<-as.matrix(d1)
   t1<-sum(d1)
   d2<-d1-t1     #complemento al totale
   t2<-sum(d2)
   p<-d2/t2
   index<-dix[1:k]
   index<-as.matrix(index)


   matr<-xc[index,y]


   matr<-as.matrix(matr)
   med<-t(matr)%*%p        #calcolo della media pesata
con pesi
   med<-as.matrix(med)     # inversamente proporzionali
alla distanza

   #sostituzione dei valori stimati ai valori mancanti
   for (i in 1:ncol(xm)){
       if(is.na(xm[t,i])){
         h<-h+1
         xm2[t,i]<-med[h,1]
       }
   }
}
dati<-riordina(xm2,xc,mat)
}
```

A2.3.r7

```
#mette i geni della matrice xc ed xm nello stesso
ordine in cui erano originariamente nel dataset

riordina<-function (da,xc,mat)
{
x<-rep(0,ncol(xc))
j<-0
k<-0
for(i in 1:nrow(mat)){
```

```
   if (mat[i,]==0){
      j<-j+1
      x<-rbind(x,xc[j,])
      }
   if(mat[i,]!=0){
      k<-k+1
      x<-rbind(x,da[k,])
      }
   }
x<-x[-1,]
rownames(x)<-c(1:nrow(mat))
x
}
```

A2.3.r8 `#calcolo del RMS error`

```
perimput.na<-function (sim,vera,nna)
{
cfr<-vera-sim
cfr<-cfr**2
s<-sum(cfr)
RMS<-s/nna
RMS
}
```

A2.3.r9 `#data la matrice xc, simula dai valori mancanti`

```
sost.na<-function (dat.sim,xc)

{
data.sim<-xc
for(i in 1:nrow(xc)){
    for(j in 1:ncol(xc)){
       if (is.na(dat.sim[i,j])){
          data.sim[i,j]<- dat.sim[i,j]
          }
       }
   }
data.sim
}
```

A2.3.r10 `#conta quanti sono i valori mancanti per ciascuna unità`

```
conta<-function (data)
{
count<-rep(0,nrow(data))
for (i in 1:nrow(data)){
    for (j in 1:ncol(data)){
        if (is.na(data[i,j])){
           count[i]<-count[i]+1
        }
```

```
    }
}
count
}
```

A2.3.r11 `#conta quanti sono i valori mancanti per ciascun gene`

```
contag<-function (data)
{
count<-rep(0,ncol(data))
for (i in 1:ncol(data)){
    for (j in 1:nrow(data)){
        if (is.na(data[j,i])){
           count[i]<-count[i]+1}
        }
    }
count
}
```

A2.3.r12 `#elenca i geni che hanno x valori mancanti`

```
pergene.na<-function (mat,x)      #; mat è l'output
della funzone contag

{
y<-0
z<-0
count<-0
for (i in 1:nrow(mat)){
    y<-y+1
    if (mat[i,1]==x) {
       z<-z+1
       count[z]<-y
    }
}
count
}
```

A2.3.r13 `#plotta i dati mancanti`

```
plot.na<-function (data)
{
x<-0
plot(1,1,ylim=c(1,nrow(data)),xlim=c(1,ncol(data)),type=
"n",main="Disposizione dei valori
mancanti",xlab="gene",ylab="Soggetto")
for (i in 1:nrow(data)){
    y<-0
    x<-x+1
    abline(h=x)
    for (j in 1:ncol(data)){
        y<-y+1
        if (is.na(data[i,j])){
```

```
            points(y,x,pch=3)
          }
#else{
#points(y,x,col=2,pch=3)}
    }
}
}
```

Capitolo 3
Analisi esplorativa

3.1 Introduzione

Nel seguito, verranno illustrate alcune analisi condotte al fine di esplorare le potenzialità di alcune tecniche di analisi multivariata nell'estrapolare informazioni d'interesse biologico sui geni disponibili nel *dataset*. Con *informazioni di interesse biologico* si intendono informazioni sia sul livello di espressione anomala (sovraespressione/sottoespressione) di alcuni singoli geni sia sulla co-espressione di gruppi di geni specifici di una certa patologia (ossia sregolati solo in una certa patologia). Si è pertanto fatto il primo tentativo di isolare geni che presentano caratteristiche differenti da una forma di leucemia all'altra.

3.2 Curve di Andrews

Il primo passo di analisi esplorativa è stato mosso mediante le curve di Andrews (si veda D. F. Andrews rif [58]). Questo metodo trasforma ogni osservazione in una serie di Fourier troncata, ovvero mappa un osservazione p-dimensionale $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ in uno spazio bidimensionale con una funzione del tipo

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2 sen\, t + x_3 \cos t + x_4 sen\, 2t + x_5 \cos 2t \ldots$$

che viene poi riportata con una rappresentazione grafica sul dominio $[-p, p]$.

Ciò che rende questo strumento uno dei più apprezzati per l'analisi esplorativa sono alcune proprietà di cui gode, desiderabili soprattutto per lo studio di dati multivariati. Essa infatti:

(1) *Preserva la me*dia.

Se si indica con $\overline{x}$ il vettore delle medie di n osservazioni p-variate x_i, allora la funzione relativa alla media $\overline{x}$, coincide con la media delle funzioni corrispondenti alle n osservazioni. In formule si ha

$$f_{\overline{x}}(t) = \frac{1}{n}\sum_{i=1}^{n} f_{x_i}(t)$$

(2) *Preserva le distanze*.

Definendo la distanza tra due funzioni come:

$$\| f_x(t) - f_y(t) \|_{L_2} = \int_{-p}^{p} \left[f_x(t) - f_y(t) \right]^2 dt$$

si ha una proporzionalità tra la distanza che separa la due funzioni calcolate in x ed y e la distanza Euclidea tra gli stessi punti, ossia

$$\| f_x(t) - f_y(t) \|_{L_2} = p \, \| x - y \|^2 = p \sum_{i=1}^{p} (x_i - y_i)^2$$

Grazie a questa relazione, punti nello spazio p-dimensionale tra loro vicini (lontani) saranno trasformati in funzioni altrettanto vicine (lontane) nello spazio bidimensionale. Allo stesso modo, gruppo di osservazioni affini o tra loro correlate (*cluster*) oppure osservazioni anomale saranno ben identificabili visivamente nel grafico della funzione. È proprio questa caratteristica di preservare le distanze che determina il coefficiente $1/\sqrt{2}$ e restringe i coefficienti di t all'insieme degli interi.

(3) *Preserva le relazioni lineari di ordinamento*

Se un punto y è collocato sulla linea che unisce x e z, allora per tutti i valori di t, $f_y(t)$ rimane tra $f_x(t)$ e $f_z(t)$.

(4) *Preserva la varianza*.

Se le osservazioni sono incorrelate e con varianza comune s^2, allora il valore della funzione in t, $f_x(t)$ ha varianza espressa da

$$\text{var}[\, f_x(t)] = s^2\left(\frac{1}{2} + sen^2\, t + \cos^2 t + sen^2\, 2t + \cos^2\, 2t \ldots\right)$$

Ora, se p è pari, la varianza si riduce ad una costante, $\frac{1}{2}s^2 p$; se invece p è dispari, la varianza varia tra $s^2(p-1)$ e $s^2(p+1)$. Si noti che nel primo caso la varianza non dipende da t e nel secondo, l'influenza di t si fa più debole all'aumentare di p. In questo modo la variabilità della funzione è quasi costante su tutto il l'intervallo $[-p, p]$, cosa che facilita l'interpretazione del grafico.

(5) *Fornisce una proiezione nello spazio unidimensionale*.

Per un particolare valore di $t = t_0$ il valore della funzione è proporzionale alla lunghezza della proiezione del vettore $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ sul vettore

$$f_1(t_0) = \left(1/\sqrt{2},\, sen\, t_0,\, \cos t_0,\, sen\, 2t_0,\, \cos\, 2t_0, \ldots\right)$$

Poiché

$$f_x(t_0) = \left\{\frac{x'\, f_1(t_0)}{[f_1'(t_0)\; f_1(t_0)]}\right\} \times [f_1'(t_0)\; f_1(t_0)]$$

Anche il questo caso, la proiezione nello spazio unidimensionale può rivelare *cluster* di osservazioni, *outliers* od altre peculiarità che diventano evidenti in questo spazio ma, in altre circostanze risultano oscurate dalle altre dimensioni.

La prima domanda a cui rispondere è se le curve di Andrews siano in grado di evidenziare geni con comportamenti anomali. Per fornire una prima idea al riguardo, si propongono i grafici delle figure 3.1, 3.2 e 3.3. che riportano le

curve di Andrews per un gene rispettivamente molto sovraespresso, molto sottoespresso e poco espresso utilizzando le espressioni delle 22 unità.

Dal momento che questi grafici sono sensibili all'ordine con cui vengono inserite le unità statistiche, per ciascuno di questi tre geni si presentano quattro curve che corrispondono a quattro diverse permutazioni dei dati. Si noti l'evidente differenza tra il codominio delle curve corrispondenti a geni sovra e sotto espressi rispetto a quelle del mediamente espresso. I grafici sono stati prodotti mediante l'ausilio della funzione `andrews.plot`.

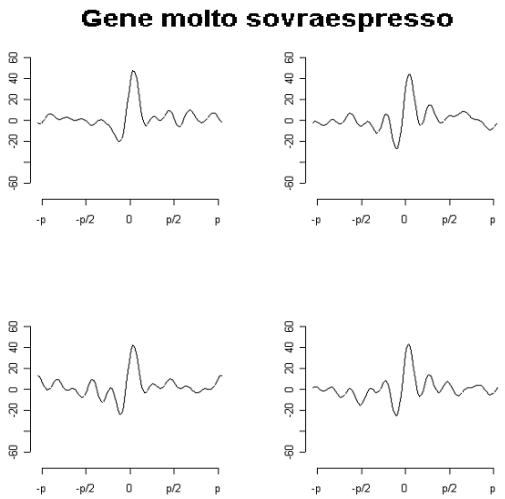

Figura 3.1

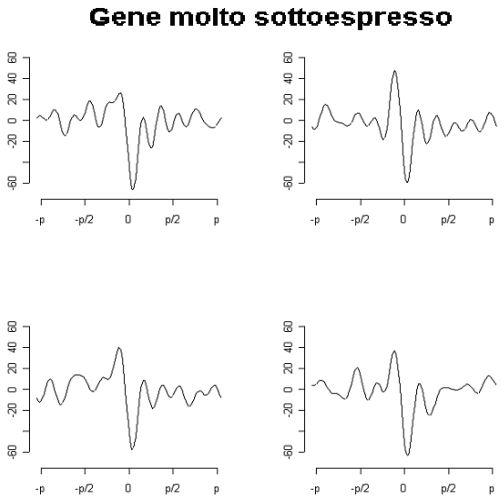

Figura 3.2

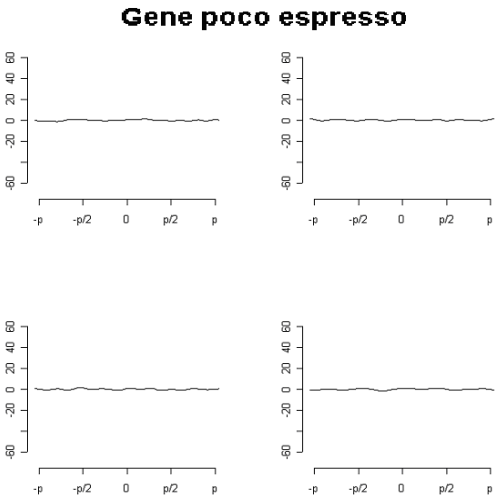

Figura 3.3

Figura 3.1 Curve di Andrews per un gene molto soveaespresso. I quattro grafici si riferiscono a quattro diverse permutazioni dei dati.

Figura 3.2 Curve di Andrews per un gene molto sottoespresso. I quattro grafici si riferiscono a quattro diverse permutazioni dei dati.

Figura 3.3 Curve di Andrews per un gene poco espresso. I quattro grafici si riferiscono a quattro diverse permutazioni dei dati.

Le figure precedenti mettono in luce la capacità delle curve di Andrews di evidenziare geni molto sovra o sotto espressi.

Una prima possibilità per evincere informazioni sui geni potrebbe quindi essere quella di costruire le curve di Andrews per tutti i geni disponibili. Sorge a questo punto il problema della dimensionalità. Infatti, dovendo analizzare graficamente un numero così elevato di variabili (4992 geni in totale), è inevitabile che i grafici risultino molto confusi e che non si riesca a risalire al comportamento di una curva per uno specifico gene al fine fare un confronto tra le patologie. È inoltre verosimile immaginare che molte informazioni riguardo ai geni differenzialmente espressi tra una patologia e l'altra non emergano chiaramente in quanto nascoste da tutti quei geni il cui livello di espressione non varia da patologia a patologia e che costituiscono quindi rumore.

3.2.1 Curve di Andrews per il *dataset* completo

Anzitutto i profili di Andrews sono stati tracciati per tutti i geni, distinguendo per ciascuna patologia, in modo da verificare se esiste una differenza di fondo tra l'espressione degli stessi geni in gruppi diversi. I grafici della Figura 3.4 mettono in luce una effettiva diversità tra i profili di gruppi diversi: in particolare risultano ben distinti i gruppi delle leucemie di tipo ALL-B e ALL-T dai gruppi ALL-Tr e AML. Si noti anche una significativa differenza tra il campo di variazione delle curve nelle diverse patologie.

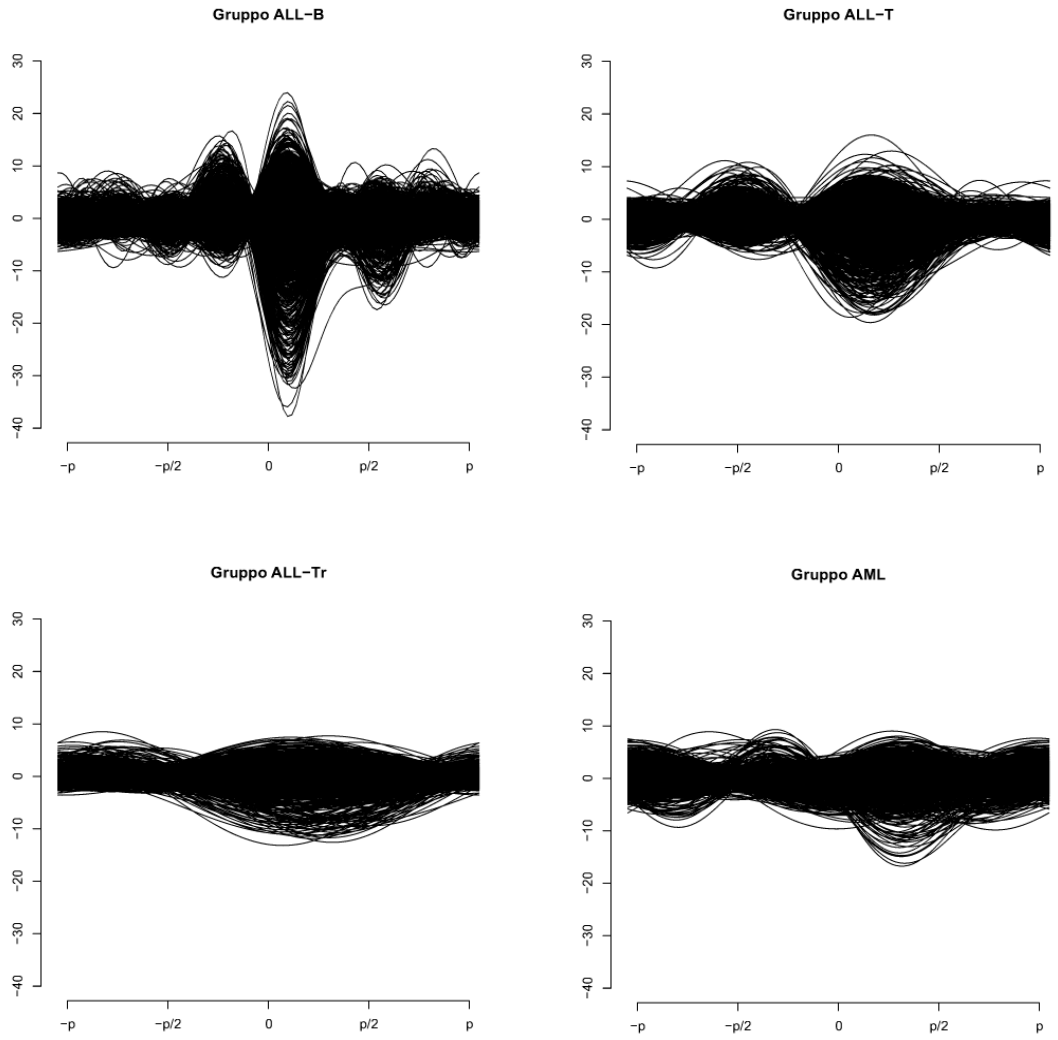

Figura 3.4 Curve di Andrews per i 4992 geni distinti nelle quattro forme di leucemia. Ciascun grafico raccoglie i profili dei geni calcolati sui solo soggetti affetti dalla corrispondente patologia.

Bisogna comunque tenere presente che la forma delle curve di Andrews dipende anche dal numero di osservazioni di cui si dispone. Dal momento che i gruppi hanno numerosità diverse, è possibile che le (a volte notevoli) discrepanze tra le curve siano imputabili a questo più che ad una effettiva differenza dei livelli d'espressione dei geni.

Per verificare questa ipotesi si sono selezionati tre soggetti a caso in ciascun gruppo in modo da avere la stessa numerosità in ogni classe e si sono nuovamente tracciati i profili, proposti in Figura 3.5. Effettivamente ora le differenze si sono appianate, ma continuano comunque a sussistere. In tutti i

casi è difficile individuare diversità tra due curve relative allo steso gene a causa dell'elevata dimensionalità del *dataset* che probabilmente occulta le differenze tra le espressioni del singolo gene.

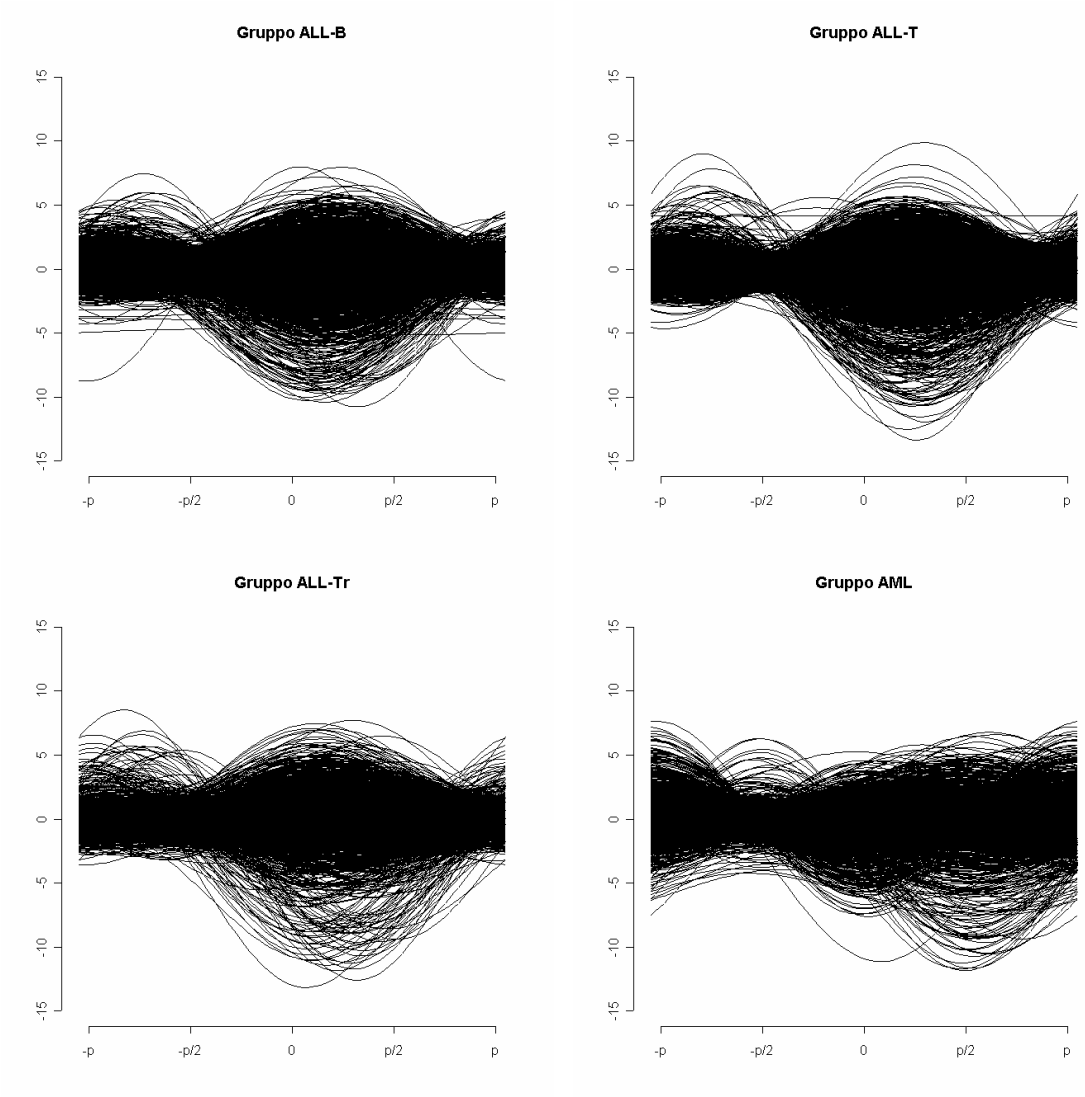

Figura 3.5 Curve di Andrews per i 4992 geni distinti nelle quattro forme di leucemia. Ciascun grafico raccoglie i profili dei geni calcolati per soli tre soggetti affetti dalla corrispondente patologia.

La funzione `andrews.plot`, con l'ausilio della quale si sono tracciati i profili di Andrews, fornisce in *output* l'elenco dei geni le cui curve oltrepassano due bande fissate in modo da lasciar uscire il 25% dei profili sia verso l'alto che verso il basso.

Per ciascun gruppo si sono dunque messi a confronto gli elenchi di questi geni (attraverso la *routine* A3.5.r1) per avere una prima idea su quali di essi hanno comportamenti diversi a seconda della patologia. Un secondo problema è quello di estrapolare informazioni specifiche per una fissata patologia. Questo può fornire informazioni sull'efficacia discriminante. È possibile infatti che un gene abbia un buon potere discriminante se il suo profilo di Andrews fuoriesce (o non fuoriesce) dalle bande per un solo tipo di patologia.

Il confronto è stato condotto fissando le bande a diversi livelli, in particolare oltre al 25% si è usato il 20%, 15%, 10%, 5%, 3%, 1%. In appendice si forniscono i risultati di questa prima analisi (Tabella A3.8 e A3.9). Al fine di rendere visibili a colpo d'occhio i diversi comportamenti dei geni nei quattro gruppi, si propongono i grafici della Figura 3.6 in cui vengono evidenziati con un punto pieno i geni le cui curve di Andrews escono dalle bande. La collocazione dei punti (ossia dei geni) all'interno del singolo grafico non ha particolare rilevanza se non per il fatto che, nei vari confronti tra grafici, a posizione uguale corrisponde lo stesso gene. La colorazione si fa invece più intensa man mano che le due bande si separano ossia è più intensa per quei geni le cui curve raggiungono livelli più alti o più bassi nell'asse delle ordinate. Le immagini sono state realizzate con le *routines* A3.5.r2 e A3.5.r3.

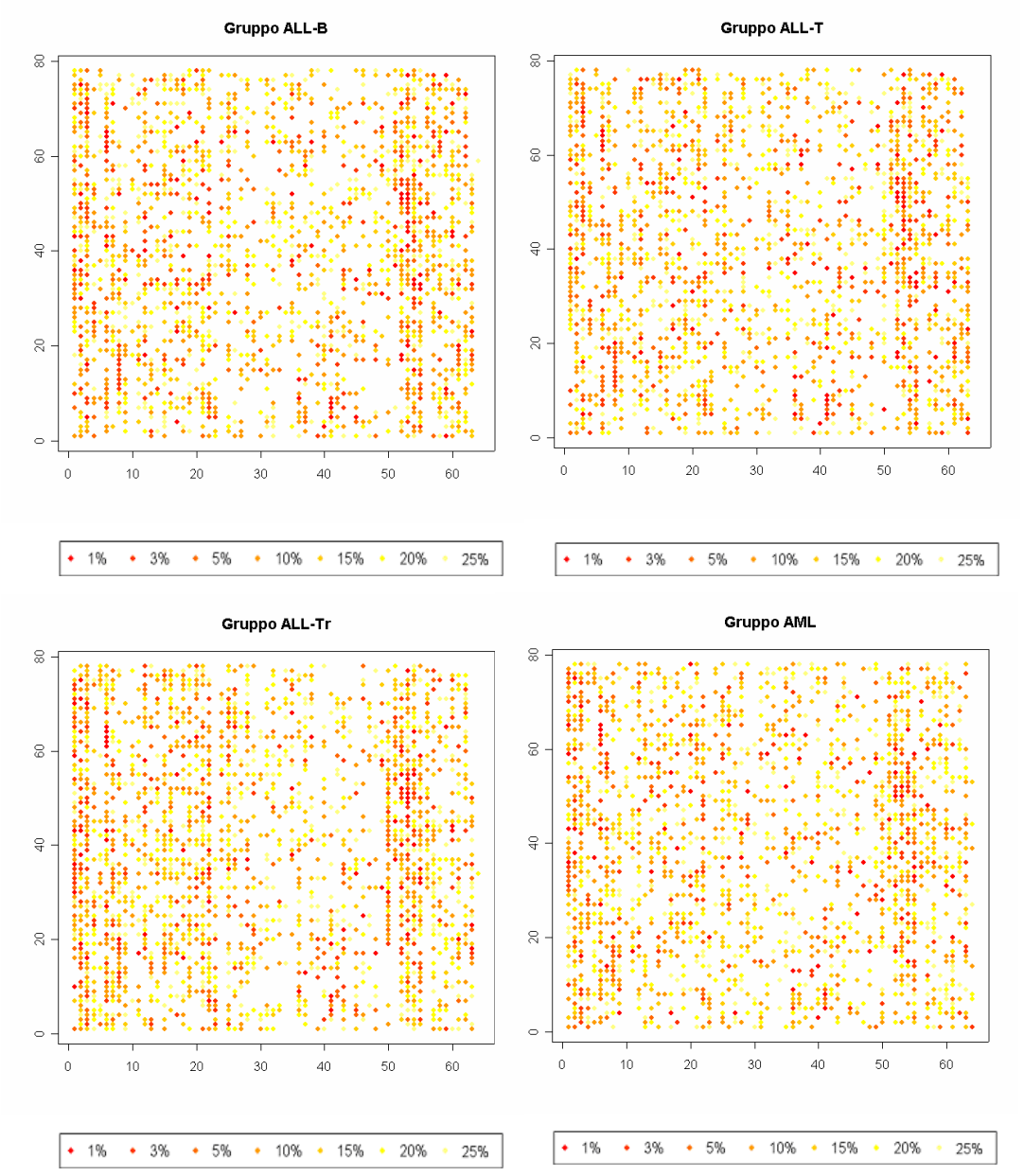

Figura 3.6 Confronto tra patologie: ciascun grafico rappresenta i geni le cui curve di Andrews fuoriescono dalle bande fissate a diverse percentuali. La colorazione si fa più intensa per quei geni che generano curve di Andrews che raggiungono livelli più alti o più bassi nell'asse delle ordinate. La disposizione dei punti all'interno della stessa matrice non ha particolare rilevanza, ma permette di confrontare le varie patologie: infatti, ogni gene occupa la medesima posizione in tutti quattro grafici.

Complessivamente si può affermare che la collocazione dei geni che fuoriescono dalle bande non varia molto da gruppo a gruppo, ossia i geni che presentano curve di Andrews con un campo di variazione più ampio in un gruppo, in linea di massima presentano i campi di variazione maggiori anche negli altri gruppi o, al contrario, i geni che hanno un andamento regolare in un gruppo, godono della stessa caratteristica anche negli altri, è probabile che si tratti di geni la cui espressione viene alterata da tutte le forme di leucemia.

In un secondo momento queste bande sono state fissate non più con criteri statistici ma partendo da ± 4 sull'asse delle ordinate, fino ad arrivare a ± 30 con un passo di 2, in modo da avere una visione non più relativa ma assoluta. Ora, l'analisi dei grafici in Figura 3.1 mette in luce delle notevoli differenze, soprattutto per quanto riguarda il primo ed il secondo gruppo: ciò significa che le curve di Andrews degli stessi geni all'interno di gruppi diversi, presentano andamenti dissimili ed hanno campo di variazione differenti.

In realtà, un'analisi di questo tipo dovrebbe essere condotta su tutte le possibili permutazioni delle unità prese tre a tre da ciascun gruppo. Una volta costruite, per ogni tripletta, le liste dei geni che fuoriescono dalle varie bande, queste dovrebbero essere poste a confronto all'interno di ciascun gruppo per vedere quali geni generano curve che escono dalle bande per tutte le triplette. A questo punto si dovrebbero selezionare quei geni le cui curve di Andrews escono (o rimangono all'interno) della bande per un solo gruppo. In questo elaborato non è stata affrontata un'analisi di questo tipo in quanto avrebbe richiesto tempi di elaborazione troppo lunghi. Forse un'analisi del genere avrebbe realmente messo in luce geni dotati di un buon potere discriminante o geni rilevanti dal punto di vista biologico.

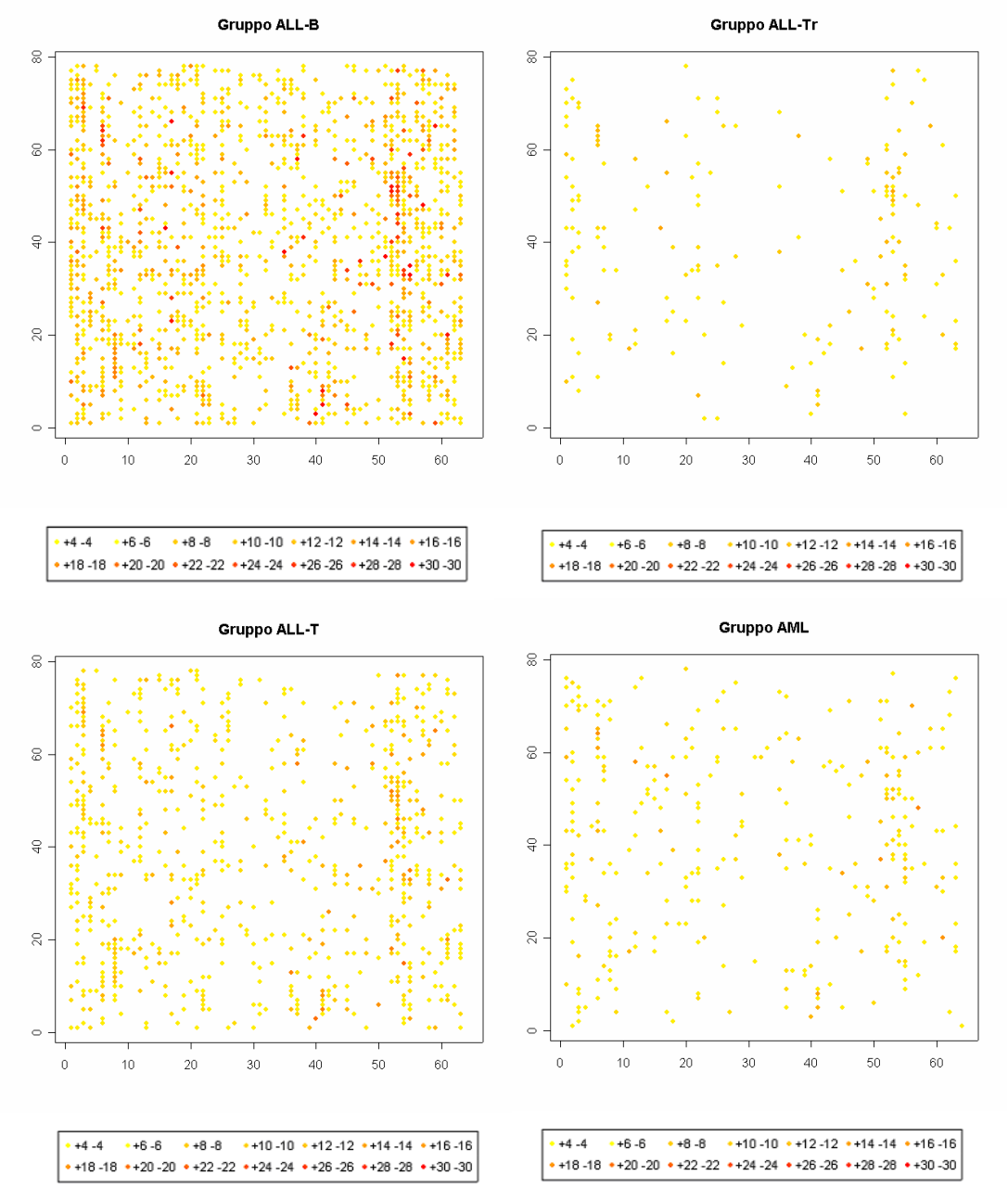

Figura 3.7 Confronto tra patologie: ciascun grafico rappresenta i geni le cui curve di Andrews fuoriescono dalle bande fissate a vari livelli sull'asse delle ordinate, partendo da ±4 fino ad arrivare a ±30 con passo di due; i geni con i profili che toccano livelli più alti o più bassi nell'asse delle ordinate, vengono evidenziati con colori man mano più intensi Anche in questo caso, i grafici non vanno considerati singolarmente ma devono essere letti in un'ottica di confronto.

3.2.2 Curve di Andrews sul *dataset* ridotto

Le stesse analisi illustrate nel paragrafo precedente sono state condotte anche sui 218 geni selezionati dai ricercatori del CRIBI in quanto geni di una certa rilevanza.

I quattro grafici della Figura 3.8 rappresentano i profili delle espressioni dei geni per ciascuna patologia. L'andamento riflette i risultati ottenuti con l'analisi su tutti i 4992 geni.

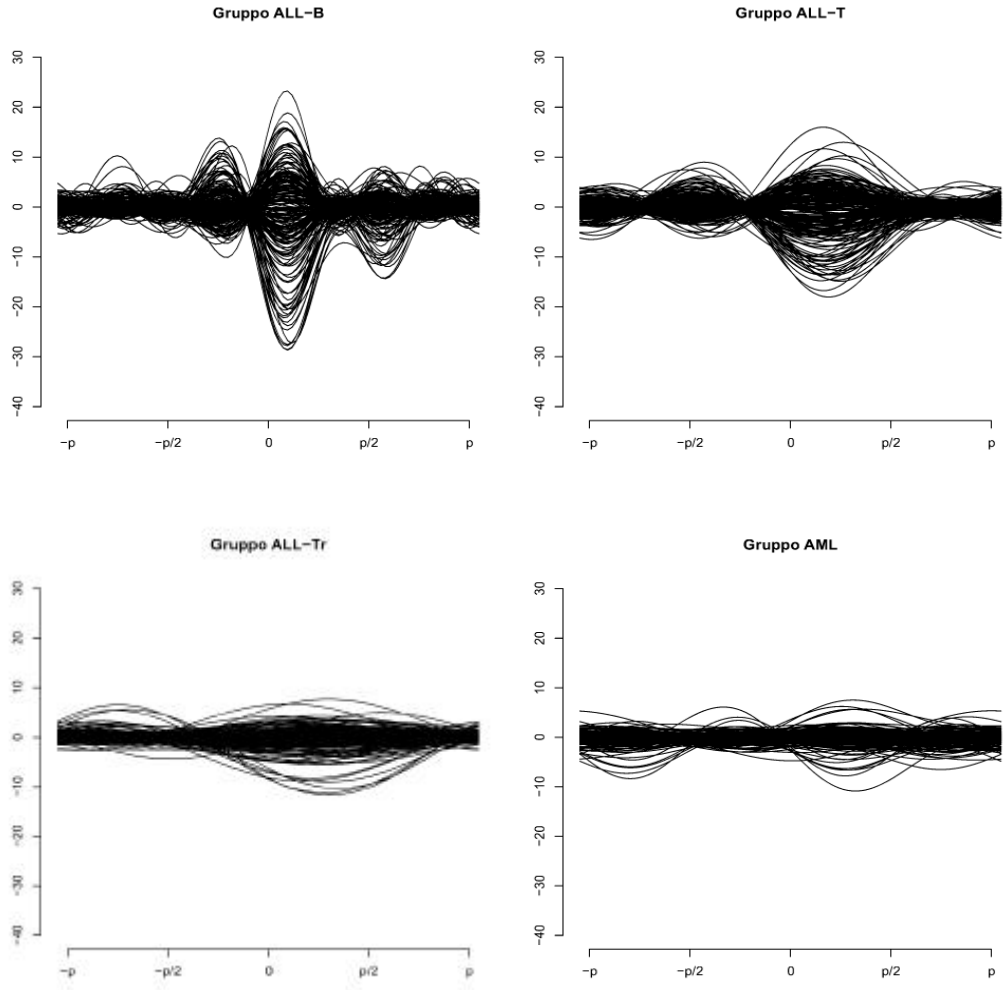

Figura 3.8 Curve di Andrews dei 218 geni della selezione distinti nelle quattro forme di leucemia; ciascun grafico raccoglie i profili dei geni calcolati sui solo soggetti affetti dalla corrispondente patologia.

La Figura 3.9 riporta le collocazioni dei geni che generano profili che escono dalle bande fissate a diverse percentuali (essa è stata realizzata con le *routines* A3.5.r4-A3.5.r5): risulta ora evidente che le curve che escono da queste bande sono generate grossomodo dagli stessi geni in tutti quattro i gruppi, a variare è invece l'intensità della colorazione, ossia il livello toccato dalle curve.

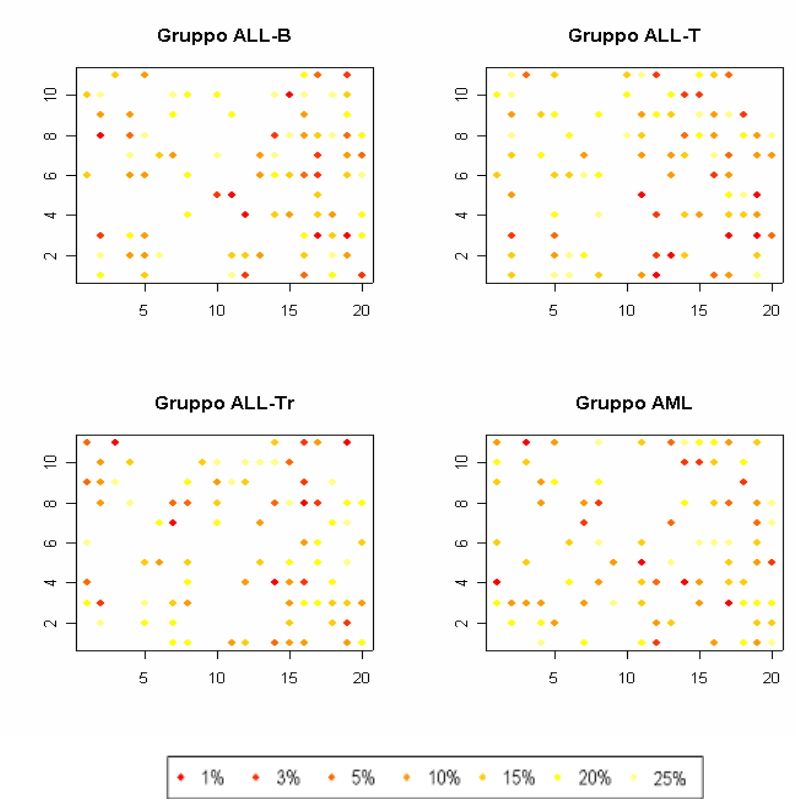

Figura 3.9 Confronto tra patologie con l'utilizzo dei soli 218 geni della selezione. La colorazione dei punti si fa più intensa per quei geni che generano curve di Andrews che raggiungono livelli più alti o più bassi nell'asse delle ordinate. La disposizione dei punti all'interno della stessa matrice non ha particolare rilevanza, ma permette di confrontare le varie patologie infatti ogni gene occupa la medesima posizione in tutti quattro grafici.

Questo si nota anche nella Figura 3.10 in cui vengono riportati i geni le cui curve superano varie soglie da ±4 a ±30; in particolare i geni del gruppo ALL-B sono quelli che generano curve che toccano i livelli più alti, verosimilmente

sarà pertanto più facile discriminare questo gruppo dagli altri. Per quanto riguarda i gruppi ALL-Tr e AML, si noti che nel primo vengono evidenziati più geni, ma quelli che vengono evidenziati nel secondo sono gli stessi che compaiono nel primo, pertanto è probabile che quei geni che vengono evidenziati nel gruppo ALL-Tr ma non in AML abbiamo un buon potere discriminante.

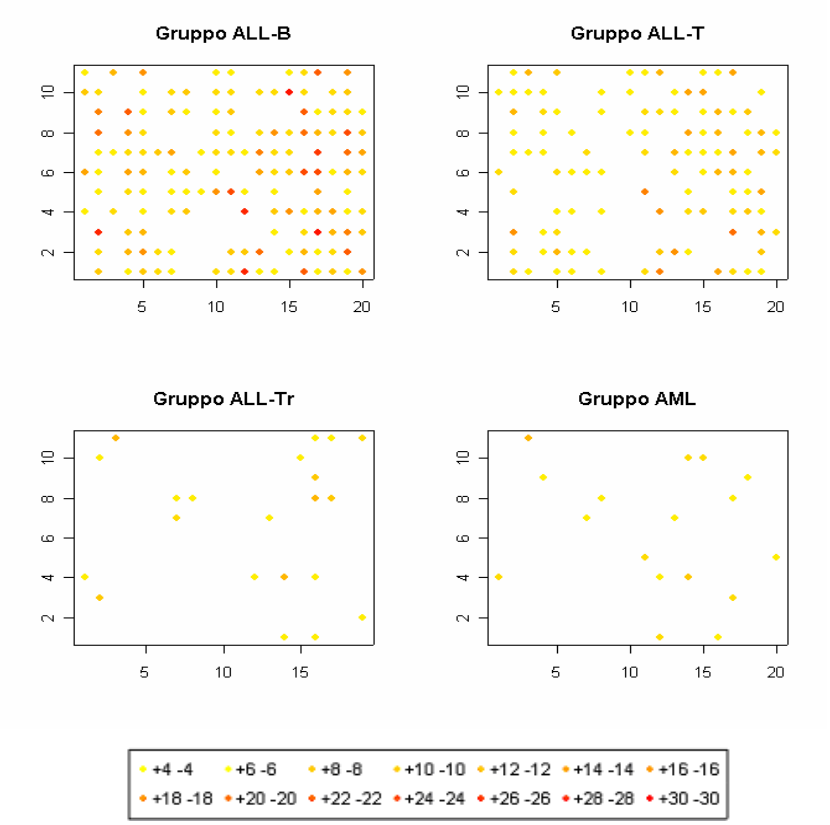

Figura 3.10 Confronto tra patologie con l'utilizzo dei soli 218 geni della selezione ciascun grafico rappresenta i geni le cui curve di Andrews fuoriescono dalle bande fissate a vari livelli sull'asse delle ordinate, partendo da ±4 fino ad arrivare a ±30 con passo di due; i geni con i profili che toccano livelli più alti o più bassi nell'asse delle ordinate, vengono evidenziati con colori man mano più intensi Anche in questo caso i grafici non vanno considerati singolarmente ma devono essere letti in un'ottica di confronto.

In appendice (Figura A3.1 e A3.2) si propongono i diagrammi a scatola di alcuni dei geni le cui curve fuoriescono (o non fuoriescono) dalle bande per un

solo tipo di patologia. Confrontando le distribuzioni sembra che alcuni di questi siano effettivamente dotati di un buon potere discriminante, almeno con riguardo ad una singola patologia. È altresì vero che altri generano dei diagrammi a scatola (relativi alle quattro patologie) che tendono a sovrapporsi ed a non evidenziare differenze nella distribuzione. In definitiva, è possibile che questo metodo sia in grado di evidenziare geni differenzialmente espressi, anche se è evidente che in alcuni casi seleziona geni che non sembrano dotati di buon potere discriminante. Come accennato in precedenza, un' analisi più approfondita probabilmente fornirebbe evidenze più concrete e permetterebbe di valutare questo metodo di selezione in maniera più adeguata e definitiva.

3.3 Analisi di raggruppamento sui geni

Nel paragrafo precedente si è tentato, con un approccio prevalentemente di tipo grafico, di mettere in luce geni il cui livello di espressione variasse da patologia a patologia. Nel seguito ci si occuperà invece di ricercare gruppi omogenei di geni all'interno di ciascuna patologia, in modo da mettere in luce se geni facenti capo a diverse patologie vengono suddivisi ed associati tra loro in modo simile o se, invece, un gruppo di geni che per una patologia possono definirsi affini, e quindi associati al medesimo *cluster*, per un'altra vengono separati e assegnati a gruppi diversi.

Questo studio è stato condotto mediante una *analisi di raggruppamento*. L'analisi di raggruppamento è, per l'appunto, un'analisi di tipo esplorativo volta a far emergere all'interno del *dataset* gruppi di osservazioni affini od omogenei al fine di dare una prima idea su quale sia la struttura dei dati anche in vista di analisi successive. Il concetto di affinità o somiglianza viene tradotto in linguaggio statistico con la nozione di *distanza*: quanto più due osservazioni sono *vicine*, tanto più sono tra loro simili. Sono stati definiti diversi criteri per misurare la distanza tra unità, tra gruppi o tra unità e gruppi e si sono date diverse definizioni di distanza secondo il tipo di variabili di cui si dispone (qualitative, quantitative, dicotomiche, politomiche ecc.). Per un'introduzione all'analisi *cluster* ed una rassegna di distanze e metodi di misura si veda Richard C. Dubes and Anil K. Jain, (1988) rif [49] oppure L. Kaufman and P. J. Rousseeuw, (1990) rif [50].

Per tutte le analisi di cui si discuterà nel seguito, si sono adottate due distanze; quella Euclidea e quella di Manhattan. Dal momento che i due approcci hanno riportato risultati analoghi, si presenteranno solo quelli relativi alla distanza Euclidea, essendo più impiegata in letteratura, seppure le considerazioni non cambierebbero di molto adottando quella di Manhattan.

La scelta che pare più sensata è quella di formare tre, al massimo quattro *cluster* di geni per ciascuna patologia, in quanto aumentando le suddivisioni si

ottengono sempre non più di quattro *cluster* abbastanza numerosi e gli altri che raccolgono al massimo una decina di geni.

3.3.1 Analisi di raggruppamento: *dataset* completo

Per misurare la distanza tra *cluster*, si sono adottati quattro differenti legami: singolo, medio, completo e di Ward. Quello completo e quello di Ward hanno fornito i risultati più soddisfacenti. Una sintesi dei risultati viene proposta nelle Tabella 3.1 e 3.2.

VARIABILI STANDARDIZZATE		Suddivisione in tre cluster			Suddivisione in quattro cluster			
Metodo	Patologia	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Legame singolo	ALL-B	4990	1	1	4989	1	1	1
	ALL-T	4989	2	1	4988	2	1	1
	ALL-Tr	4990	1	1	4987	3	1	1
	AML	4990	1	1	4989	1	1	1
Legame medio	ALL-B	4983	8	1	4983	7	1	1
	ALL-T	4832	145	15	4832	145	14	1
	ALL-Tr	4908	83	1	4908	48	35	1
	AML	4956	35	1	4956	25	10	1
Legame completo	ALL-B	3593	1340	59	3593	1199	141	59
	ALL-T	4724	169	99	4306	418	169	99
	ALL-Tr	4009	926	57	4009	926	36	21
	AML	3701	1246	45	3701	1215	45	31
Legame di Ward	ALL-B	2586	1890	516	2586	1570	516	320
	ALL-T	2858	1260	874	2858	1260	765	109
	ALL-Tr	3253	1024	715	3253	811	715	213
	AML	3913	798	281	3913	501	297	281
Metodo delle k-medie	ALL-B	3796	856	340	3531	668	640	153
	ALL-T	3623	898	471	3422	775	658	137
	ALL-Tr	4314	491	187	3961	515	433	83
	AML	4200	526	266	3930	459	407	196
Metodo dei medoidi	ALL-B	3515	913	564	3229	913	673	177
	ALL-T	3341	989	662	3146	971	674	201
	ALL-Tr	3935	555	502	3727	570	555	140
	AML	3667	705	620	3162	786	672	372

Tabella 3.1 Numerosità dei *cluster* per le varie tecniche di raggruppamento applicate sul *dataset* completo utilizzando variabili standardizzate. La parte a sinistra propone la suddivisione in tre gruppi, quella a destra in quattro.

VARIABILI NON STANDARDIZZATE		Suddivisione in tre cluster			Suddivisione in quattro cluster			
Metodo	Patologia	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Legame singolo	ALL-B	4990	1	1	4989	1	1	1
	ALL-T	4989	2	1	4988	2	1	1
	ALL-Tr	4990	1	1	4987	3	1	1
	AML	4990	1	1	4989	1	1	1
Legame medio	ALL-B	4889	102	1	4888	102	1	1
	ALL-T	4819	166	7	4819	164	7	2
	ALL-Tr	4930	44	18	4930	29	18	15
	AML	4935	49	8	4933	49	8	2
Legame completo	ALL-B	3874	1036	82	3874	1036	76	6
	ALL-T	4296	685	11	4296	642	43	11
	ALL-Tr	4371	566	55	4185	566	186	55
	AML	4855	84	53	4855	69	53	15
Legame di Ward	ALL-B	3212	1117	663	3212	928	663	189
	ALL-T	2387	1709	896	2387	1709	788	108
	ALL-Tr	3571	902	519	3571	732	519	170
	AML	4049	517	426	3413	636	517	426
Metodo delle k-medie	ALL-B	3710	992	290	3477	686	682	147
	ALL-T	3589	943	460	3390	787	673	142
	ALL-Tr	4324	488	180	3920	557	435	80
	AML	3845	766	381	3941	471	398	182
Metodo dei medoidi	ALL-B	3490	950	552	2808	1017	895	272
	ALL-T	3314	1019	659	2143	1373	911	565
	ALL-Tr	3788	650	554	3641	652	551	148
	AML	2736	1703	553	2803	1188	544	457

Tabella 3.2 Numerosità dei *cluster* per le varie tecniche di raggruppamento applicate sul *dataset* completo utilizzando variabili non standardizzate. La parte a sinistra propone la suddivisione in tre gruppi, quella a destra in quattro.

Con il metodo del legame singolo (o del vicino più prossimo), la distanza tra due *cluster* C_1 e C_2 viene definita come:

$$D(C_1, C_2) = min\{ d(i, j); \quad i \in C_1 \, , \, j \in C_2 \}$$

e vengono pertanto aggregati i due *cluster* per cui la distanza tra le due unità più vicine è minima. Applicata al *dataset* in esame, questa strategia tende a generare un solo *cluster* numeroso e tutti gli altri con uno, due o al massimo tre geni.

Il legame medio definisce la distanza tra *cluster* come:

$$D(C_1, C_2) = \frac{\sum d(i, j)}{n_1 \, n_2}$$

dove le quantità a denominatore indicano le numerosità di ciascun *cluster*. Applicato al *dataset*, questo legame tende a creare un *cluster* molto numeroso (oltre le 4800 unità) e gli altri piuttosto piccoli (al massimo un centinaio di unità). Nel passaggio da quattro a tre gruppi viene mantenuta questa struttura.
Infine il legame completo definisce la distanza come:

$$D(C_1, C_2) = max\{d(i, j); \quad i \in C_1 , j \in C_2\}$$

e pertanto aggrega i due *cluster* per cui la distanza tra le due unità più lontane è minima. Anche in questo caso si viene a creare un grosso raggruppamento (4100-4800 unità) ma per le leucemie di tipo ALL-B si evidenzia anche un secondo gruppo abbastanza numeroso (1036 unità) e nel passaggio da quattro a tre *cluster*, vengono aggregati i gruppi meno numerosi.

Nel legame di Ward, infine, i *cluster* da aggregare sono quelli che minimizzano la devianza tra i centroidi. Pertanto la distanza tra *cluster* viene definita come:

$$D(C_1, C_2) = \left(\bar{C}_1 - \bar{C}\right)^2 + \left(\bar{C}_2 - \bar{C}\right)^2$$

dove $\bar{C}_1$ e $\bar{C}_2$ sono i centroidi dei due *cluster*, mentre $\bar{C} = \left(\bar{C}_1 - \bar{C}_2\right)/2$.

Con questo metodo, rispetto ai precedenti, si sono ottenuti *cluster* con numerosità molto più equilibrata. Tranne per le leucemie di tipo AML, si riscontra sempre un *cluster* più consistente (dalle 2300 alle 3500 unità) e gli altri con numerosità che va dalle 500 alle 1700 unità. Inoltre nel passaggio da quattro a tre gruppi, il *cluster* più corposo mantiene intatta la sua dimensione, ossia vengono aggregati i *cluster* meno numerosi. I geni che fanno capo alle leucemie AML, sembrano aggregarsi in modo diverso: se si vanno a formare tre *cluster*, uno di essi contiene oltre 4000 geni, e nel passaggio da tre a quattro, sono i gruppi meno numerosi a mantenere la loro dimensione.

Si è poi passati all'utilizzo di metodi non gerarchici di accorpamento. Sebbene i tali metodi prevedano la conoscenza del numero di gruppi in cui devono essere suddivise le unità, si è deciso, sulla base degli esiti delle analisi precedenti, di usare il metodo delle *k*-medie e dei medoidi anche per la suddivisione dei geni in tre o quattro gruppi. Anche questi risultati sono

riassunti nelle Tabella 3.1 e 3.2 e confermano l'analisi precedente: sembra sensata una suddivisione in tre o quattro gruppi ed anche in questo caso viene a crearsi un gruppo piuttosto consistente (oltre le 3000 unità). Nelle tabelle 3.3 e 3.4 si riportano i medoidi ottenuto usando le variabili standardizzate e non; si sono evidenziati quelli che non sono cambiati passando da un metodo all'altro, mentre si sono sottolineati quelli che rimangono tali passando da quattro a tre gruppi. Si noti che alcuni medoidi cambiano da patologia a patologia e nel passaggio da quattro a tre gruppi, le classi ALL-B e ALL-Tr mantengono due dei tre medoidi, le altre due classi ne mantengono soltanto uno.

La stessa analisi è stata condotta su variabili standardizzate e non. I risultati sono leggermente cambiati ma le considerazioni che si possono trarre sono sempre le stesse. In particolare, tra i metodi gerarchici, i più adeguati sembrano essere ancora il legame completo e quello di Ward; per quanto riguarda i metodi non gerarchici c'è da sottolineare che alcuni dei geni che erano stati selezionati come medoidi, ora non lo sono più, inoltre, passando da quattro a tre gruppi, viene conservato al massimo un medoide per patologia.

Geni selezionati come medoidi (variabili standardizzate)							
Patologia	Analisi con tre gruppi			Analisi con quattro gruppi			
ALL-B	1541	4935	256	1541	4935	3334	1113
ALL-T	4851	4955	4221	4851	4976	3978	541
ALL-Tr	94	1962	3646	94	1962	2119	1614
AML	331	297	4066	3966	3328	4066	2952

Tabella 3.3

Geni selezionati come medoidi (variabili non standardizzate)							
Patologia	Analisi con tre gruppi			Analisi con quattro gruppi			
ALL-B	1541	4935	256	1541	3451	3683	1129
ALL-T	4851	4955	4211	1851	2847	2019	4363
ALL-Tr	94	825	4252	94	632	3848	1614
AML	1868	2528	2407	3927	2971	915	2407

Tabella 3.4

Tabella 3.3 e Tabella 3.4 Tabelle riassuntive dei geni selezionati come medoidi (*dataset* completo) per le quattro forme di leucemia facendo uso di variabili standardizzate (Tabella 2.5) e non (Tabella 2.6). La parte a sinistra si riferisce alla suddivisione in tre gruppi, quella a destra in quattro gruppi. Sono stati sottolineati i medoidi comuni alle due suddivisioni ed evidenziati quelli selezionati sia usando variabili standardizzate che non standardizzate.

3.3.2 Analisi di raggruppamento: *dataset* ridotto

I metodi che hanno portato ad una migliore suddivisione nell'analisi sul dataste completo, sono stati applicati anche sui 218 geni della selezione. Nelle Tabella 3.5 e 3.6Tabella 3.6 si riportano i risultati prima e a seguito della standardizzazione. Per quanto riguarda i medoidi (Tabella 3.7 e 3.8), si noti che nessuno di quelli evidenziati nell'analisi con il *dataset* completo è stato riproposto anche con quello ridotto. Inoltre alcuni geni qui selezionati come medoidi rappresentativi di *cluster* diversi, appartengono allo stesso *cluster* nell'analisi con tutti i geni. Si riscontra una maggiore coerenza tra i risultati relativi a variabili standardizzate e non; inoltre si noti che buona parte dei medoidi vengono conservati nel passare da tre a quattro gruppi.

VARIABILI STANDARDIZZATE		Suddivisione in tre cluster			Suddivisione in quattro cluster			
Metodo	Patologia	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Legame completo	ALL-B	166	29	23	166	23	16	13
	ALL-T	172	41	5	102	70	41	5
	ALL-Tr	187	25	6	164	25	23	6
	AML	210	6	2	205	6	5	2
Legame di Ward	ALL-B	112	78	28	78	70	42	28
	ALL-T	123	71	24	71	68	55	24
	ALL-Tr	144	43	31	144	43	25	6
	AML	133	57	28	133	51	28	6
Metodo delle k-medie	ALL-B	102	92	24	96	58	43	21
	ALL-T	106	85	27	73	72	49	24
	ALL-Tr	154	38	26	140	36	36	6
	AML	147	64	7	81	72	58	7
Metodo dei medoidi	ALL-B	105	85	28	76	75	48	19
	ALL-T	78	72	68	72	60	59	27
	ALL-Tr	131	46	41	131	46	35	6
	AML	148	45	25	87	63	43	25

Tabella 3.5 Numerosità dei *cluster* per le varie tecniche di raggruppamento applicate sui 218 geni della selezione utilizzando variabili standardizzate. La parte a sinistra propone la suddivisione in tre gruppi, quella a destra in quattro.

VARIABILI NON STANDARDIZZATE		Suddivisione in tre cluster			Suddivisione in quattro cluster			
Metodo	Patologia	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Legame completo	ALL-B	119	81	18	119	68	18	13
	ALL-T	108	94	16	103	94	16	5
	ALL-Tr	187	25	6	164	25	23	6
	AML	205	7	6	205	6	4	3
Legame di Ward	ALL-B	107	63	48	107	63	36	12
	ALL-T	124	71	26	75	71	49	23
	ALL-Tr	144	43	31	144	43	25	6
	AML	141	57	20	97	57	44	20
Metodo delle k-medie	ALL-B	108	85	25	84	82	27	25
	ALL-T	106	85	27	74	71	49	24
	ALL-Tr	154	38	26	140	36	36	6
	AML	113	91	14	133	42	36	7
Metodo dei medoidi	ALL-B	108	82	28	74	64	53	27
	ALL-T	76	73	69	73	61	59	25
	ALL-Tr	128	47	43	128	47	37	6
	AML	138	47	33	85	59	46	28

Tabella 3.6 Numerosità dei *cluster* per le varie tecniche di raggruppamento applicate sui 218 geni della selezione utilizzando variabili non standardizzate. La parte a sinistra propone la suddivisione in tre gruppi, quella a destra in quattro.

Geni selezionati come medoidi (variabili standardizzate)							
Patologia	Analisi con tre gruppi			Analisi con quattro gruppi			
ALL-B	137	895	3334	2103	895	4015	4215
ALL-T	2714	4164	4021	2714	4218	4021	4167
ALL-Tr	1714	1413	3864	1714	1413	3864	4174
AML	2441	558	2034	2517	2254	558	2034

Tabella 3.7

Geni selezionati come medoidi (variabili non standardizzate)							
Patologia	Analisi con tre gruppi			Analisi con quattro gruppi			
ALL-B	137	895	3334	474	1710	3334	1113
ALL-T	2714	4164	2849	2714	4218	2849	4167
ALL-Tr	1714	1413	3864	1714	1413	3864	4174
AML	1619	558	3390	4142	2441	558	3390

Tabella 3.8

Tabella 3.7 e Tabella 3.8 Tabelle riassuntive dei geni selezionati come medoidi (*dataset* ridotto) per le quattro forme di leucemia facendo uso di variabili standardizzate (Tabella 2.7) e non (Tabella 2.8). La parte a sinistra si riferisce alla suddivisione in tre gruppi, quella a destra in quattro gruppi. Sono stati sottolineati i medoidi comuni alle due suddivisioni ed evidenziati quelli selezionati sia usando variabili standardizzate che non standardizzate

Tanto questi medoidi, quanto quelli relativi al *dataset* completo sono stati sottoposti all'attenzione dei ricercatori del CRIBI. Dalle loro analisi è emerso che si tratta di geni di cui si conosce davvero poco, per la maggior parte di essi, addirittura, non è nemmeno la funzione biologica.

3.3.3 Analisi di raggruppamamento: un confronto

Dopo aver classificato anche i 218 geni della selezione, con riferimento alle variabili standardizzate, si sono posti a confronto i risultati ottenuti classificando i geni dei *datasets* completo e ridotto. Le tabelle 3.9 (a)-(d) che si propongono di seguito, riguardano il legame completo; ciascuna di esse rappresenta una patologia e mette in evidenza quanti geni dal *dataset* ridotto siano stati accorpati nel medesimo *cluster* sia nell'analisi sul *dataset* ridotto che nell'analisi sul *dataset* completo. Leggendo la tabella per colonna si può vedere se e come i geni della selezione che appartengono allo stesso *cluster* sono stati separati quando si sono usati tutti i geni per la classificazione, leggendola per riga invece si intuisce quanto la classificazione che fa uso di tutti i geni rispecchia quelle che usa solo la selezione.

ALL-B		Cluster			
		1	2	3	4
Classificazione con tutti i geni: Cluster	1	101	0	0	23
	2	63	0	0	0
	3	2	14	0	0
	4	0	2	13	0

(a)

ALL-T		Cluster			
		1	2	3	4
Classificazione con tutti i geni: Cluster	1	0	23	0	5
	2	37	79	0	0
	3	33	0	18	0
	4	0	0	23	0

(b)

Legame completo		Classificazione con i geni della selezione			
ALL-Tr		Cluster			
		1	2	3	4
Classificazione con tutti i geni: Cluster	1	46	23	0	0
	2	118	0	0	24
	3	0	0	5	1
	4	0	0	1	0

(c)

Legame completo		Classificazione con i geni della selezione			
AML		Cluster			
		1	2	3	4
Classificazione con tutti i geni: Cluster	1	140	5	0	0
	2	62	0	1	2
	3	0	0	5	0
	4	3	0	0	0

(d)

Tabella 3.9(a)-(d) Confronto tra l'analisi *cluster* con il legame completo sul *dataset* completo (4992×22) e ridotto (218×22).Lette per colonna (riga) fanno capire come e se i geni che appartengono allo stesso *cluster* nell'analisi sul *dataset* ridotto (completo) vengano assegnati a *cluster* diversi dall'analisi sul *dataset* completo (ridotto). Ciascuna tabella corrisponde ad una particolare forma di leucemia.

In appendice vengono fornite le tabelle relative alle altre tecniche per la formazione di tre o quattro gruppi (Tabella A3.1-Tabella A3.6 (a)-(d)) oltre ai dendrogrammi, distinti per patologia, relativi al legame completo e al legame di Ward sia per il *dataset* completo sia per la selezione di 218 geni, distinguendo per variabili standardizzate e non Figura A3.19-Figura A3.50. Infine si propongono per ciascun metodo (ma solo per variabili standardizzate) dei grafici che mostrano come sono stati raggruppati i geni (Figura A3.3-Figura A3.18).

Appendice

A3.1 Curve di Andrews

Diagrammi a scatola relativi alle distribuzioni dei geni che generano curve di Andrews che fuoriescono (o non fuoriescono) per una sola patologia dalle bande a varie percentuali o fissate sull'asse delle ordinate partendo da ±4 con passo di due.

Alcuni di questi sembrano effettivamente dotati di un buon potere discriminante almeno rispetto ad una singola patologia, altri generano dei box-plot (relativi alle quattro patologie) che tendono a sovrapporsi ed a non evidenziare differenze nella distribuzione. In definitiva e possibile che questo metodo sia in grado di evidenziare geni differenzialmente espressi, anche è evidente che in alcuni casi seleziona geni che, non sembrano dotati di buon potere discriminante.

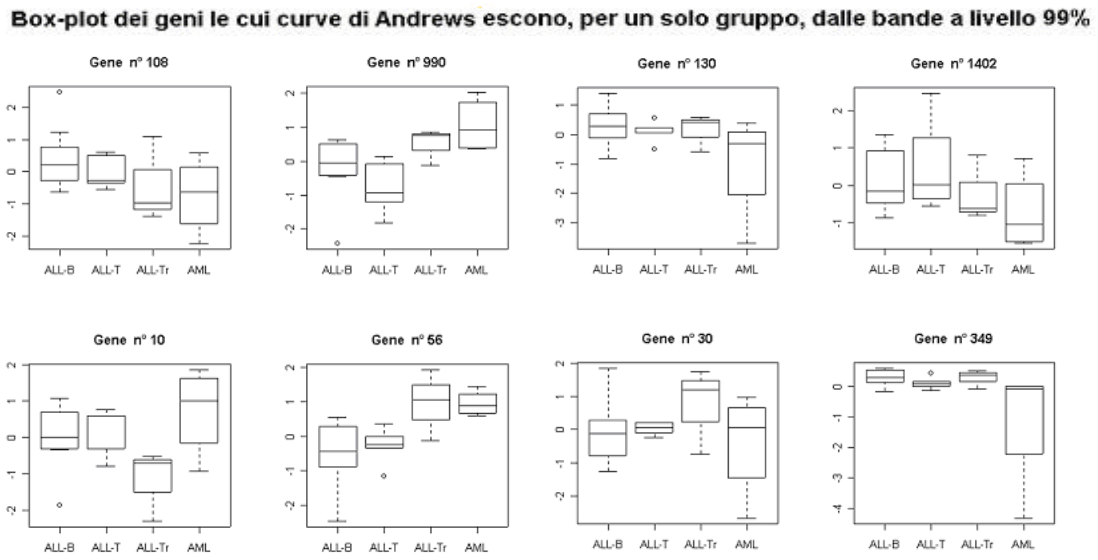

Figura A3.1 Box-plot di alcuni dei geni per i quali le curve di Andrews fuoriescono, in un solo gruppo, dalle bande fissate al 99%. Ciascun box-plot rappresenta la distribuzione dell'espressione del gene per la patologia corrispondente.

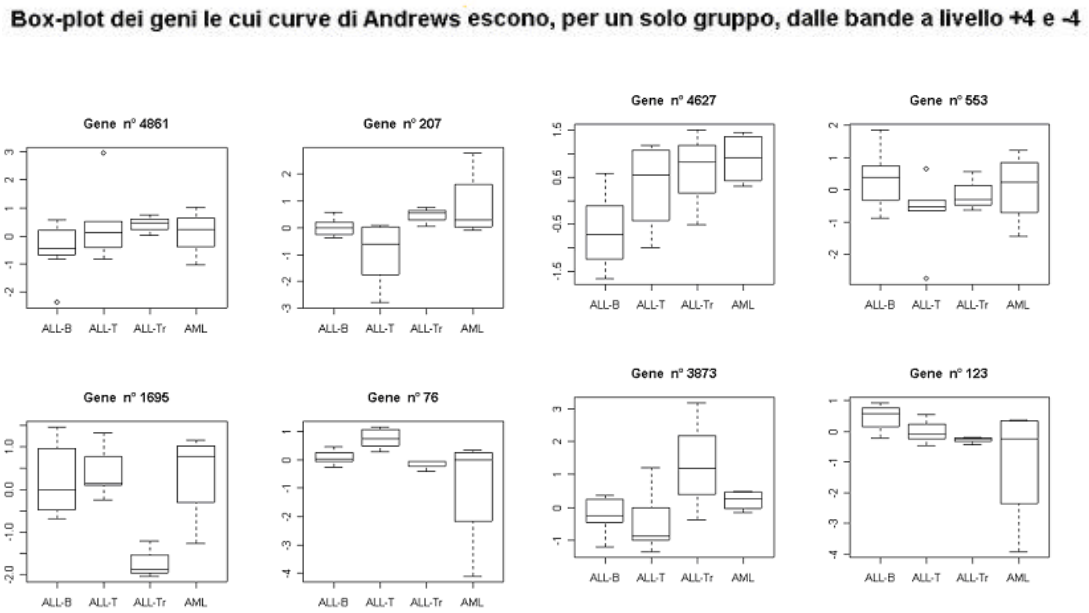

Figura A3.2 Box-plot di alcuni dei geni per i quali le curve di Andrews fuoriescono, in un solo gruppo, dalle bande fissate a livello ± 4 . Ciascun box-plot rappresenta la distribuzione dell'espressione del gene per la patologia corrispondente.

A3.2 Analisi *cluster*

Rappresentazione grafica della variazione degli esiti della classificazione per vari tipi di legame applicati al *dataset* completo e ridotto. Sull'asse verticale sono rappresentati i geni; sull'asse orizzontale le quattro patologie. Ogni segmento orizzontale rappresenta un singolo gene. La prima colonna (leucemia ALL-B) raccoglie i geni ordinati secondo il gruppo di appartenenza; nelle colonne successive gli stessi geni acquisiscono colorazioni corrispondenti alla classificazione ottenuta nella patologia. Se i geni venissero accorpati sempre nello stesso modo all'interno delle quattro patologie, le quattro colonne corrispondenti mostrerebbero la stessa colorazione. Quanto più marcata è la variazione cromatica nelle colonne successive alla prima, tanto più l'accorpamento cambia da patologia a patologia. Pertanto se il susseguirsi dei colori non differisce molto per due o più patologie, è sensato affermare che il modo con cui sono stati aggregati i geni non varia da patologia a patologia. Le Figura A3.3-Figura A3.10 riguardano la suddivisione in quattro (Figura A3.3-Figura A3.6) e in tre (Figura A3.7-Figura A3.10) *cluster* di tutti i 4992 geni del *dataset* completo, le Figura A3.11-Figura A3.18 sono invece relative al *dataset* ridotto le Figura A3.11-Figura A3.14 rappresentano la suddivisione in quattro *cluster*, le Figura A3.15-Figura A3.18 in tre. I grafici proposti sono stati realizzati nelle *routines* A3.5.r7-A3.5.r8-A3.5.r9-A3.5.r10.

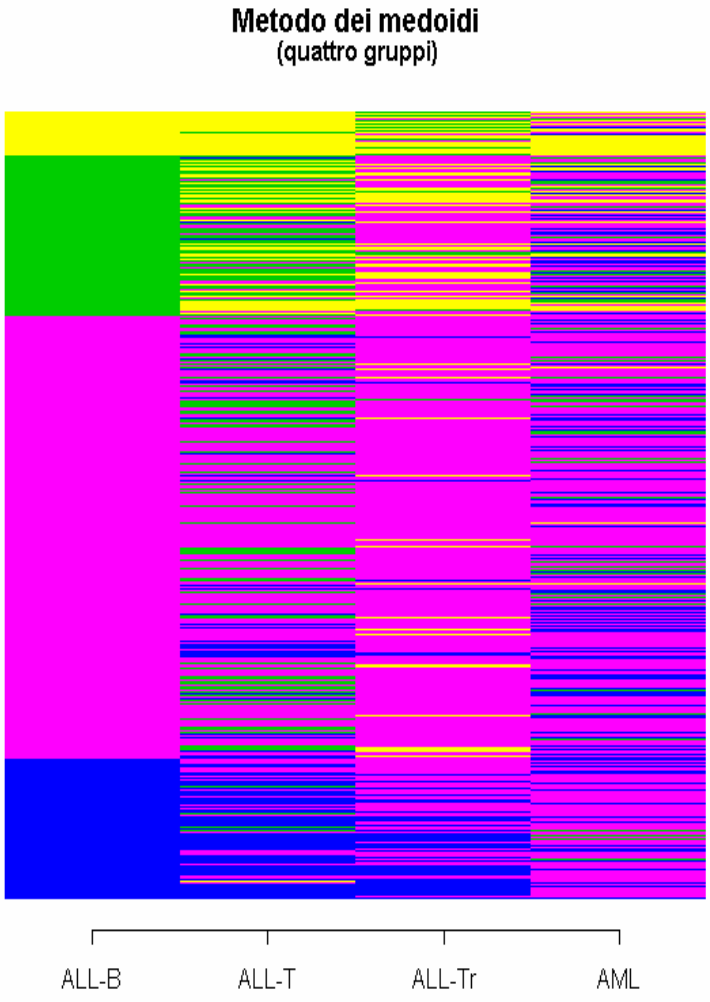

Figura A3.3 Metodo dei medoidi, suddivisione in quattro gruppi, *dataset* completo

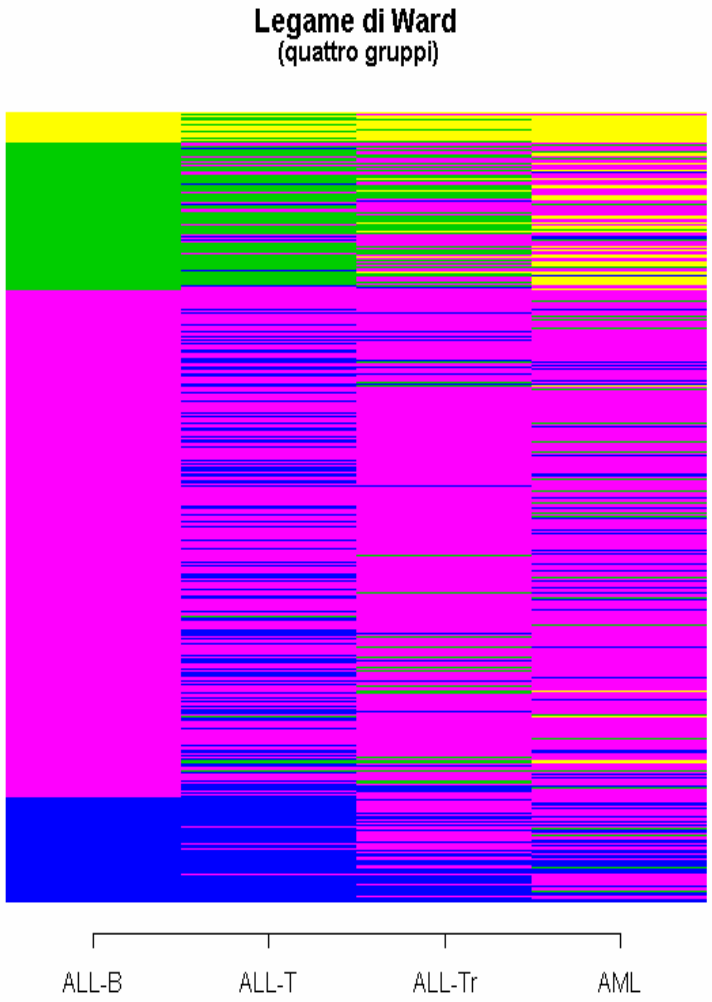

Figura A3.4 Legame di Ward, suddivisione in quattro gruppi, *dataset* completo

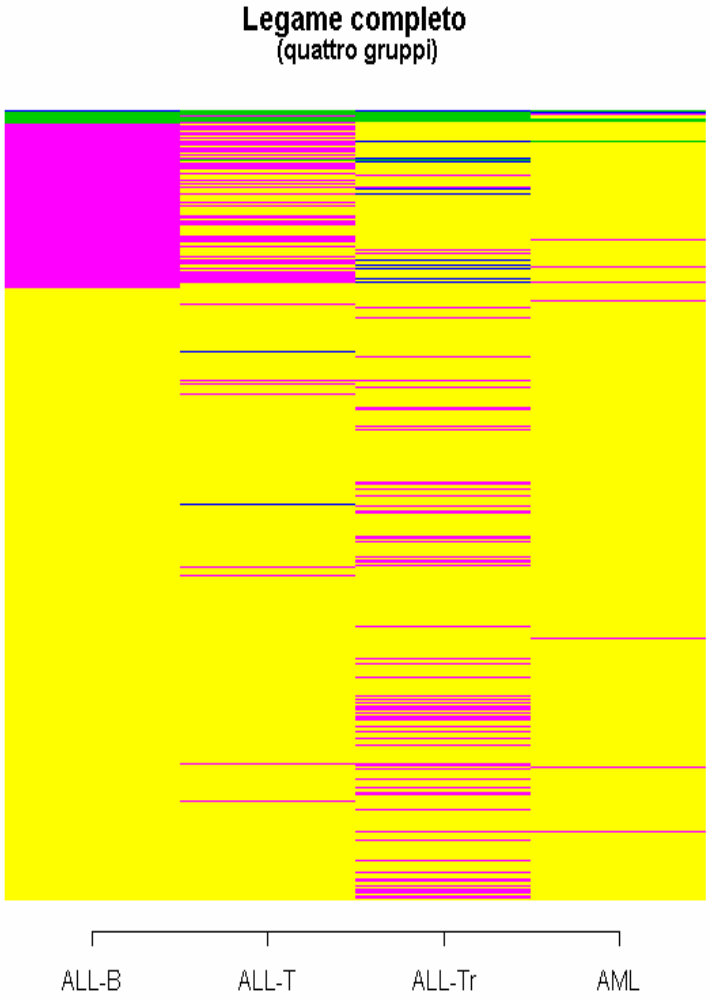

Figura A3.5 Legame completo, suddivisione in quattro gruppi, *dataset* completo

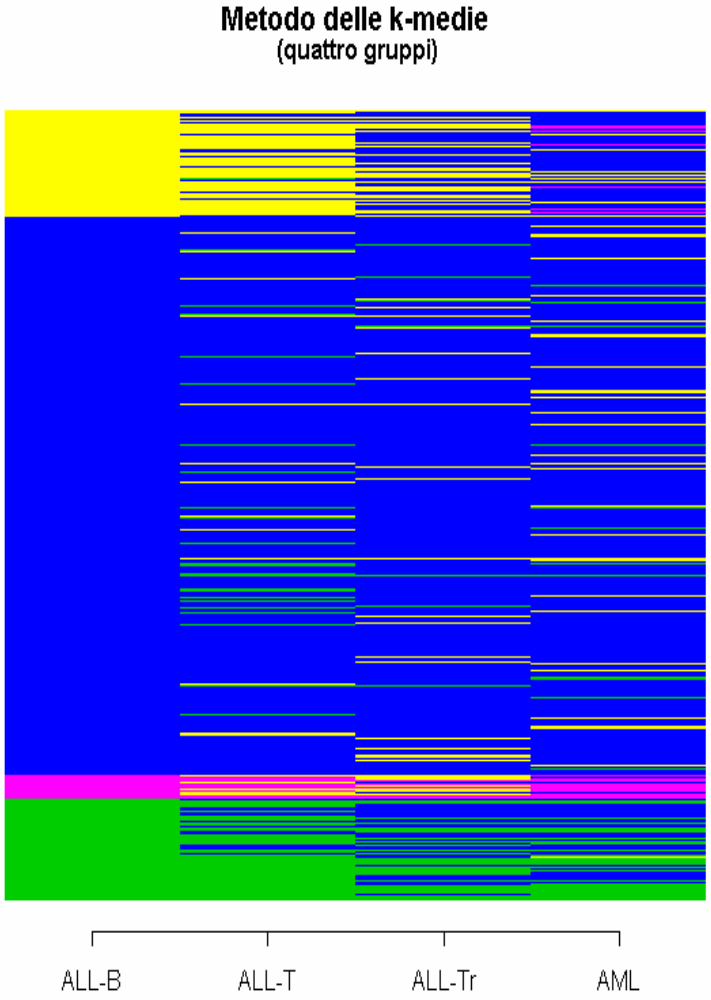

Figura A3.6 Metodo delle *k*-medie, suddivisione in quattro gruppi, *dataset* completo

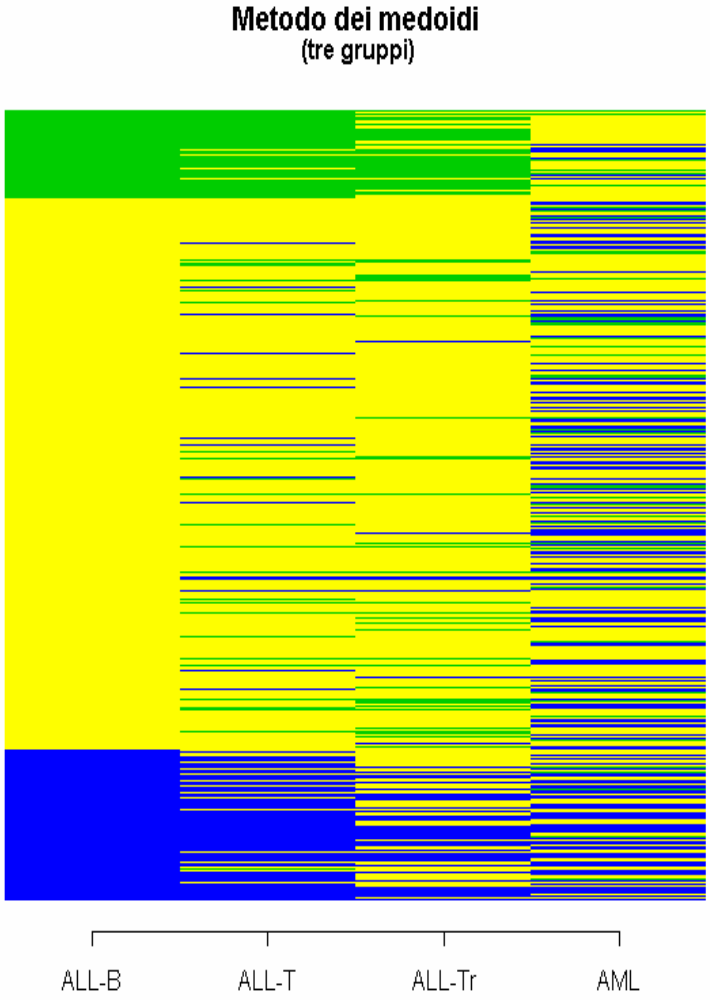

Figura A3.7 Metodo dei medoidi, suddivisione in tre gruppi, *dataset* completo

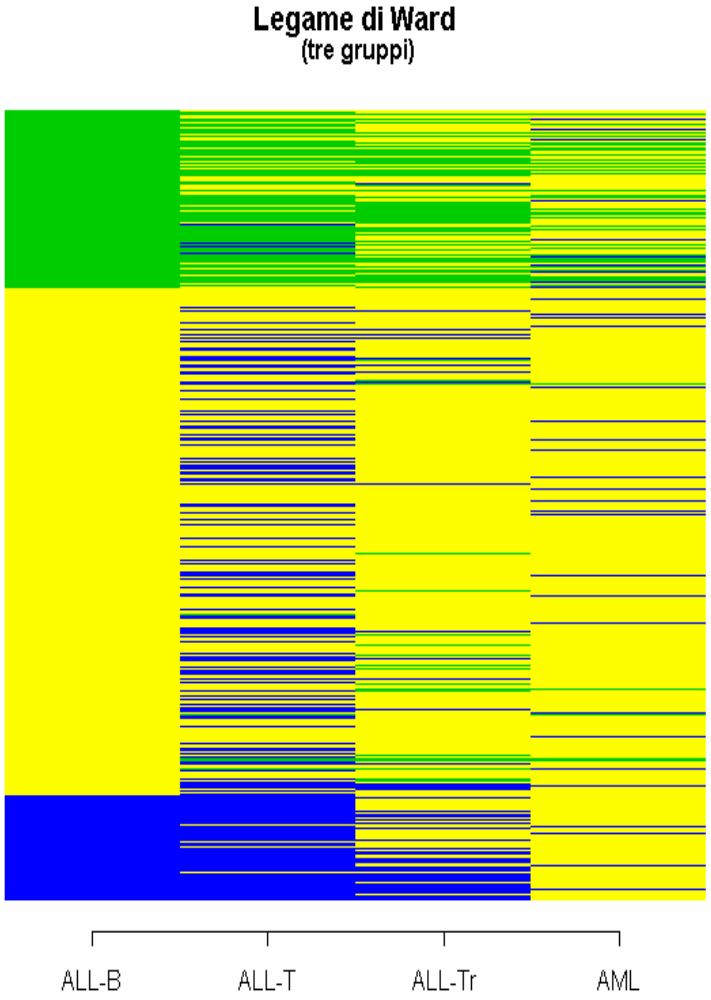

Figura A3.8 Legame di Ward, suddivisione in tre gruppi, *dataset* completo

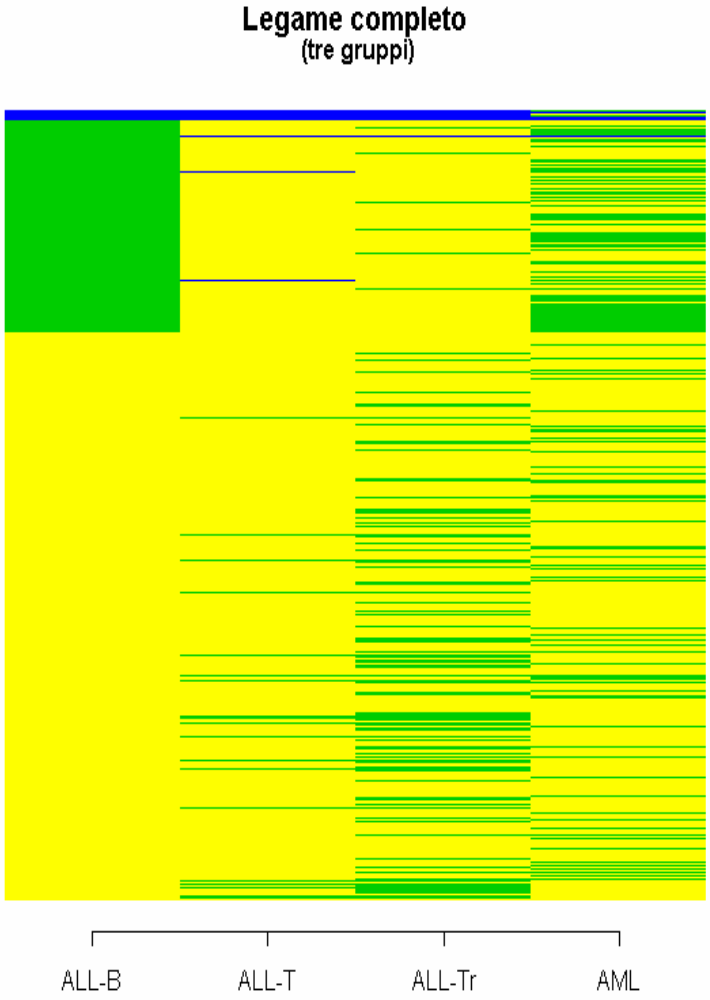

Figura A3.9 Legame completo, suddivisione in tre gruppi, *dataset* completo

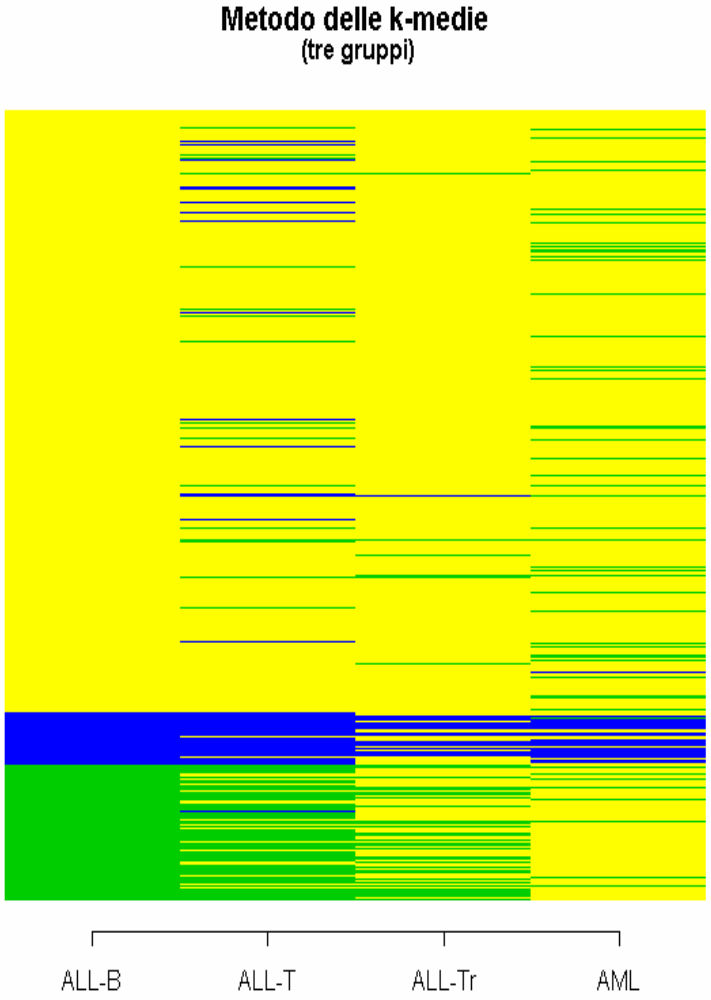

Figura A3.10 Metodo delle *k*-medie, suddivisione in tre gruppi, *dataset* completo

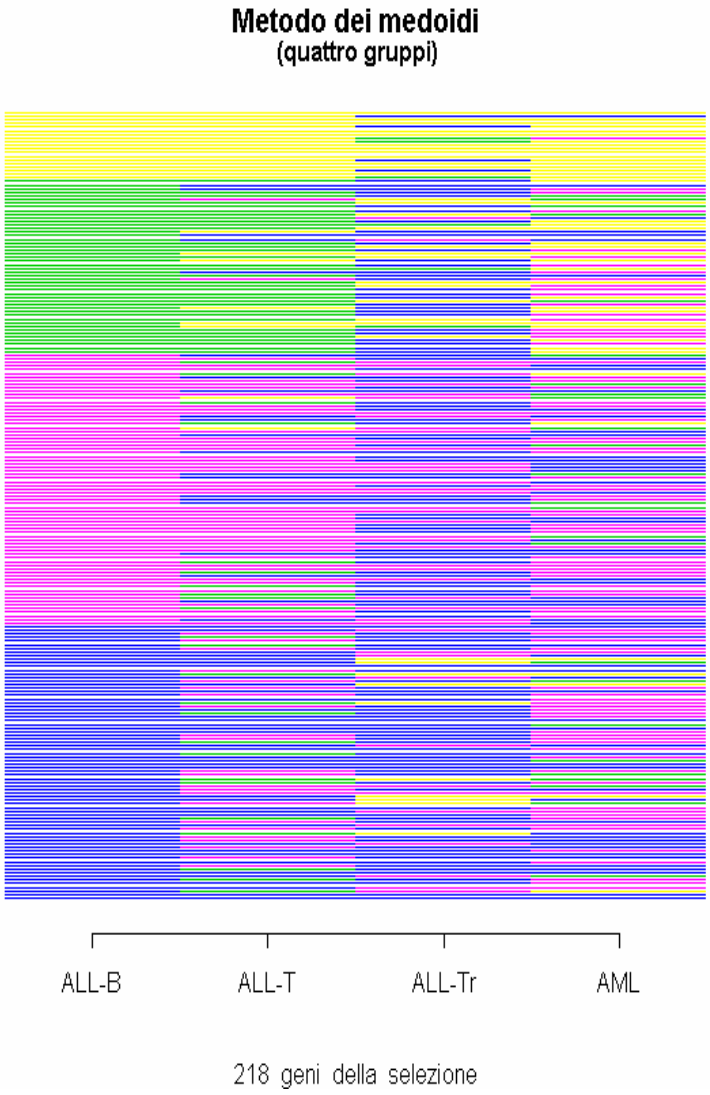

Figura A3.11 Metodo dei medoidi, suddivisione in quattro gruppi, *dataset* ridotto

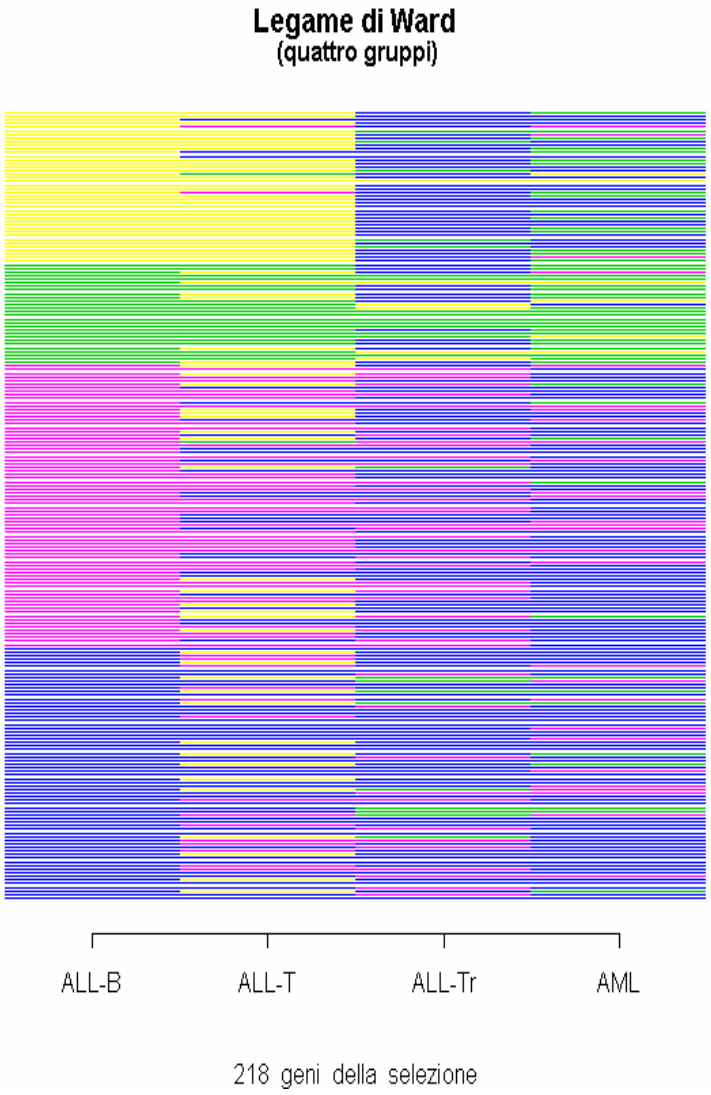

Figura A3.12 Legame di Ward, suddivisione in quattro gruppi, *dataset* ridotto

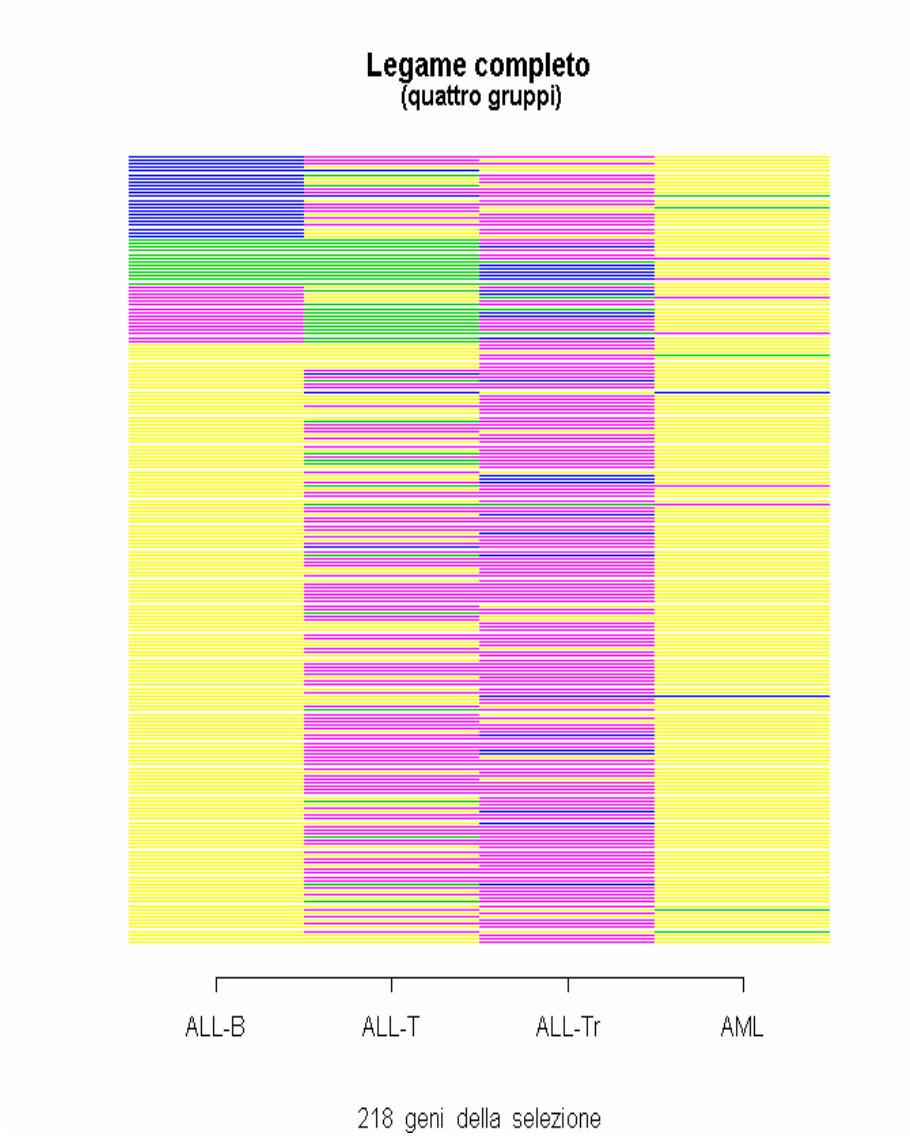

Figura A3.13 Legame di completo, suddivisione in quattro gruppi, *dataset* ridotto

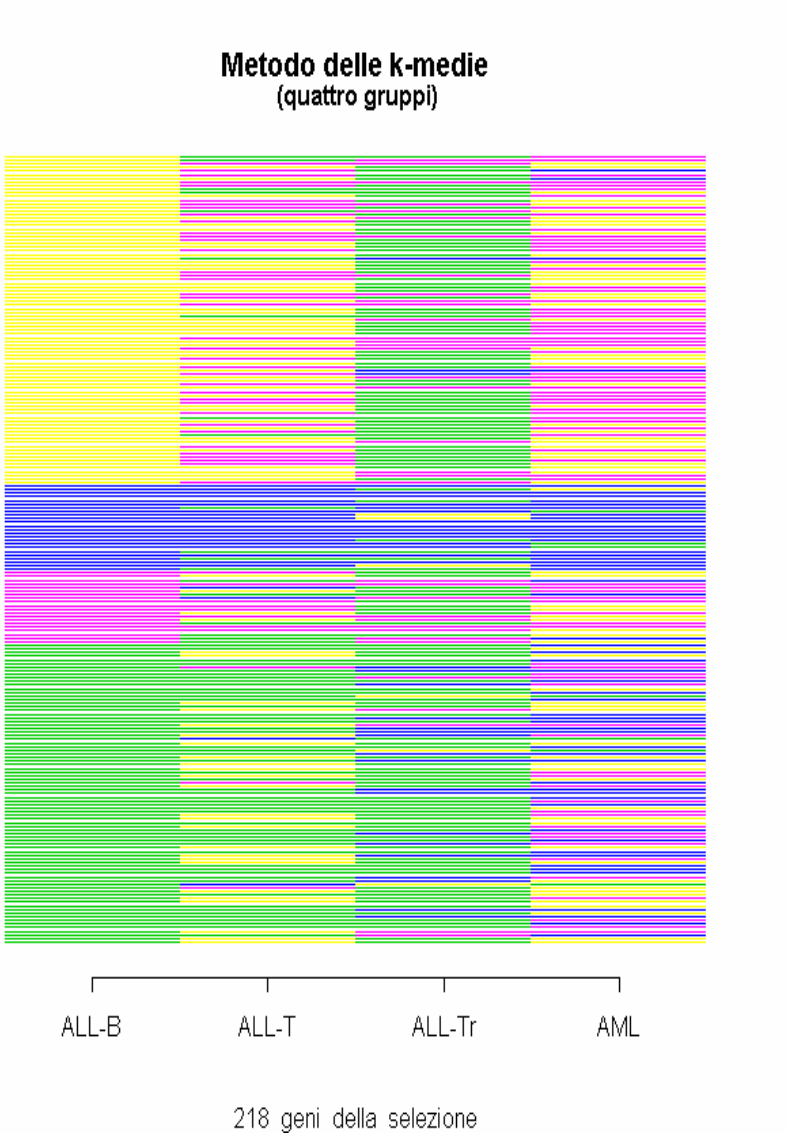

Figura A3.14 Metodo delle *k*-medie, suddivisione in quattro gruppi, *dataset* ridotto

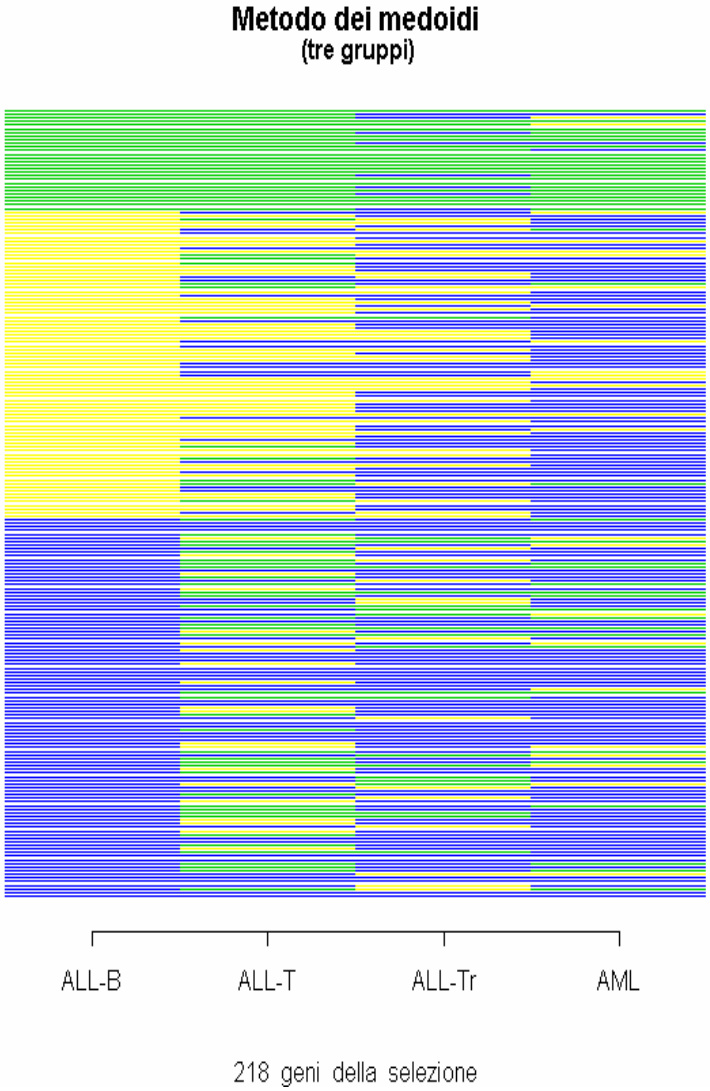

Figura A3.15 Metodo dei medoidi, suddivisione in tre gruppi, *dataset* ridotto

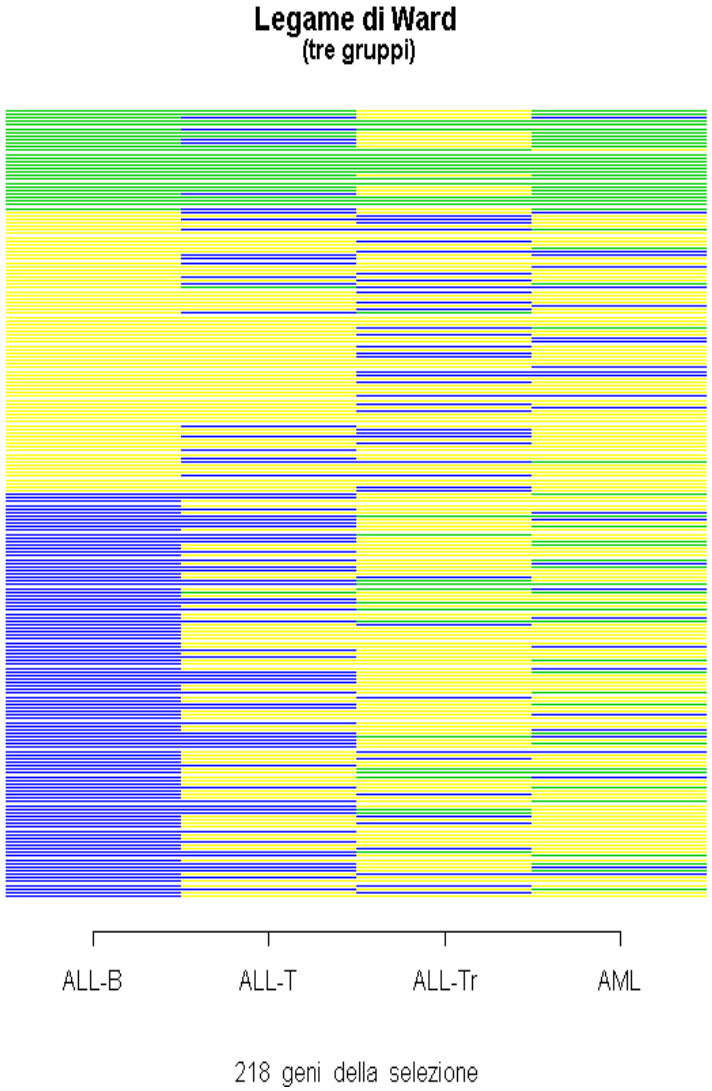

Figura A3.16 Legame di Ward, suddivisione in tre gruppi, *dataset* ridotto

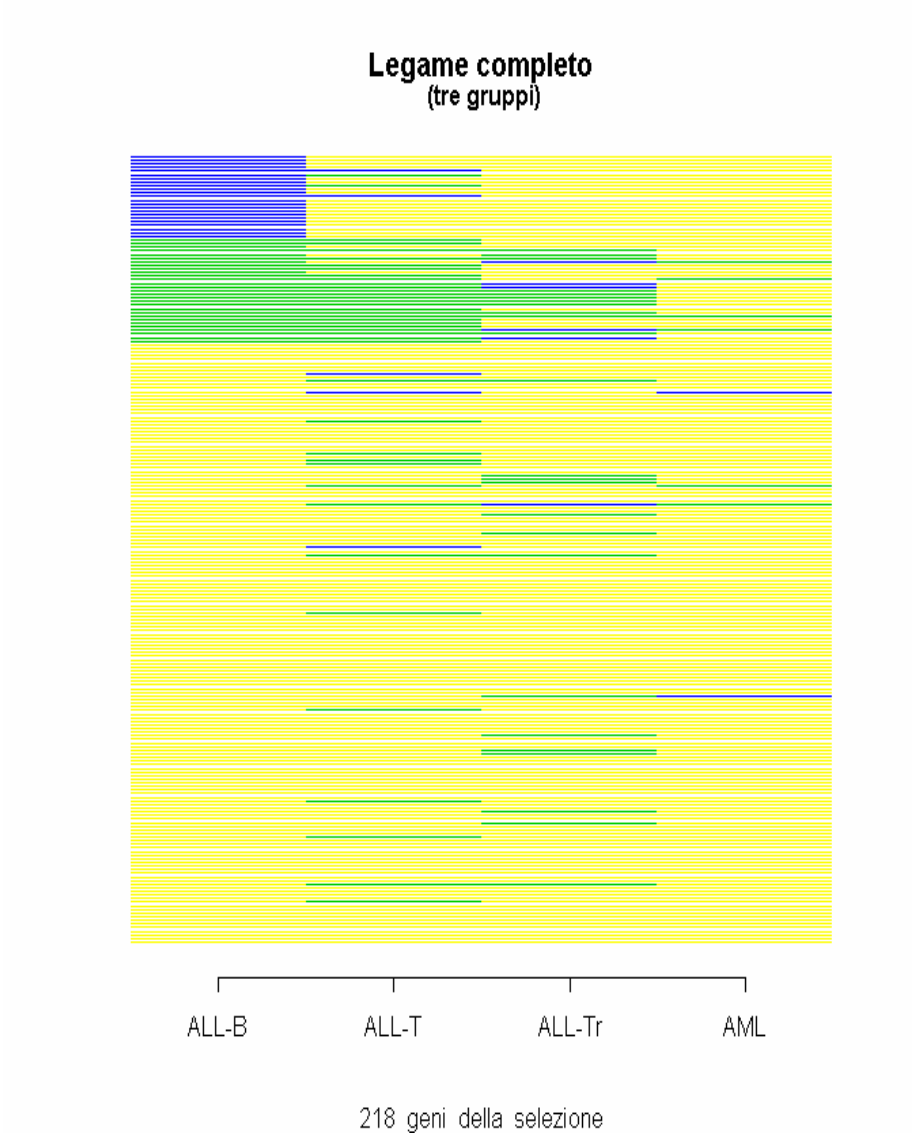

Figura A3.17 Legame completo, suddivisione in tre gruppi, *dataset* ridotto

Figura A3.18 Metodo delle *k*-medie, suddivisione in tre gruppi, *dataset* ridotto

A3.3 Analisi *cluster*: dendrogrammi

Dendrogrammi relativi ai legami completi e di Ward; vengono presentati prima quelli costruiti sul *dataset* completo (Figura A3.19-Figura A3.34) e poi quelli riguardanti il *dataset* ridotto (Figura A3.35-Figura A3.50). Inoltre, per ciascuna patologia vengono affiancati i risultati derivanti dalle applicazioni sia su variabili standardizzate che non.

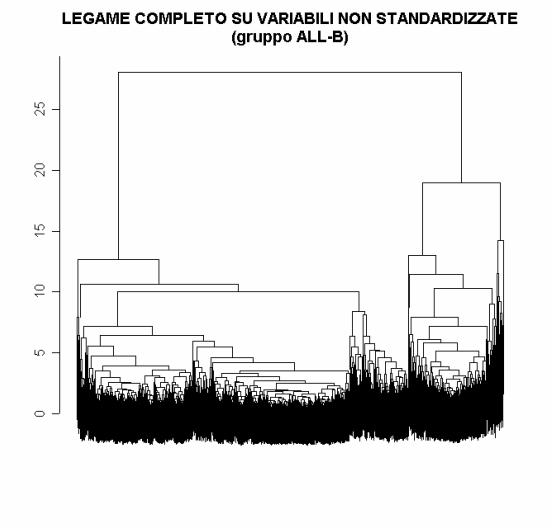

Figura A3.19 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* completo, gruppo ALL-B, variabili non standardizzate.

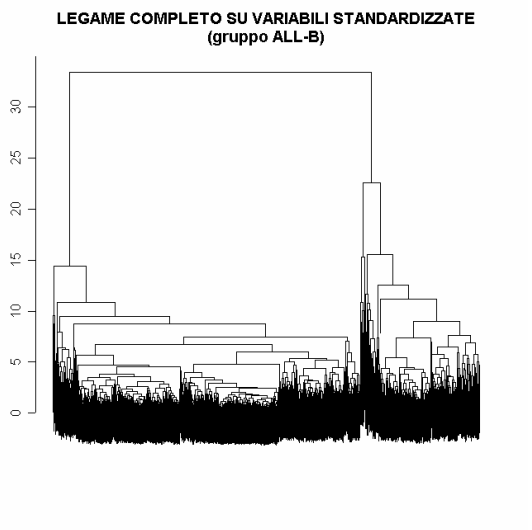

Figura A3.20 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* completo, gruppo ALL-B, variabili standardizzate.

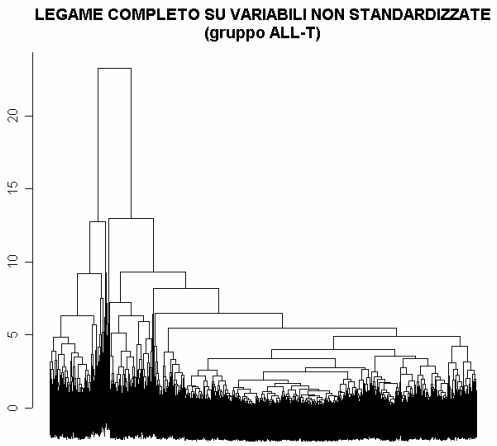

Figura A3.21 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* completo, gruppo ALL-T, variabili non standardizzate.

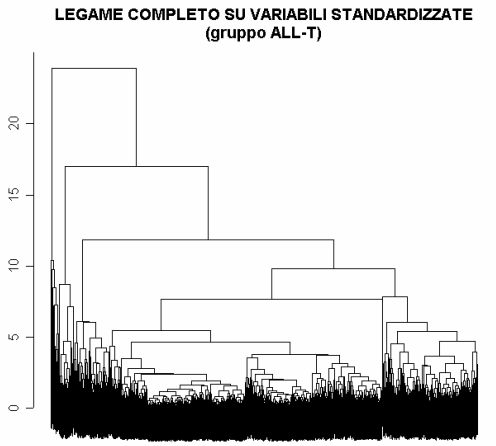

Figura A3.22 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* completo, gruppo ALL-T, variabili standardizzate.

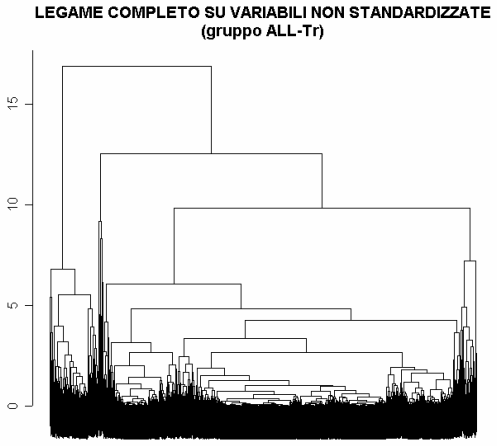

Figura A3.23 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* completo, gruppo ALL-Tr, variabili non standardizzate.

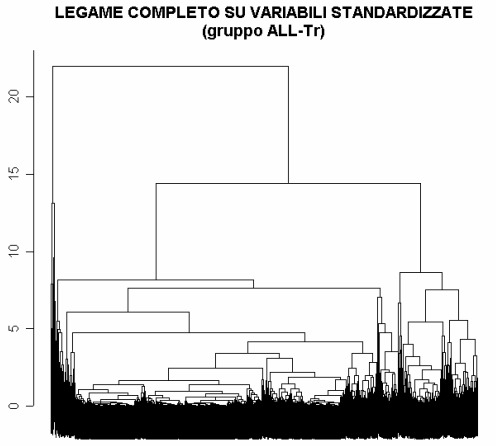

Figura A3.24 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* completo, gruppo ALL-Tr, variabili standardizzate.

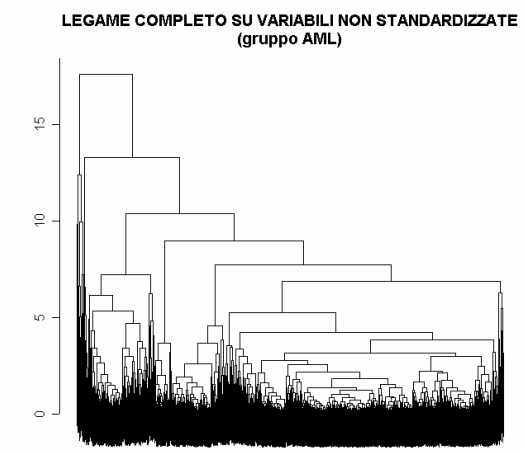

Figura A3.25 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* completo, gruppo AML,variabili non standardizzate.

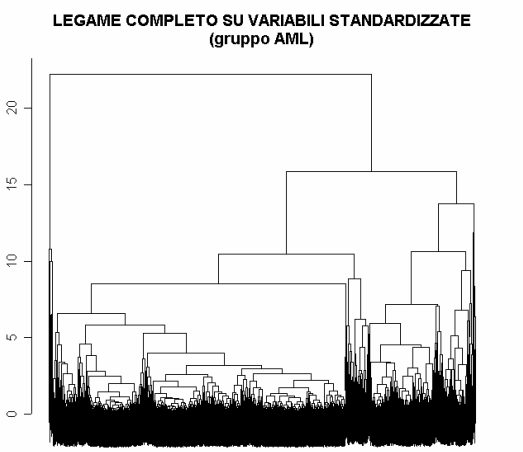

Figura A3.26 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* completo, gruppo AML,variabili standardizzate.

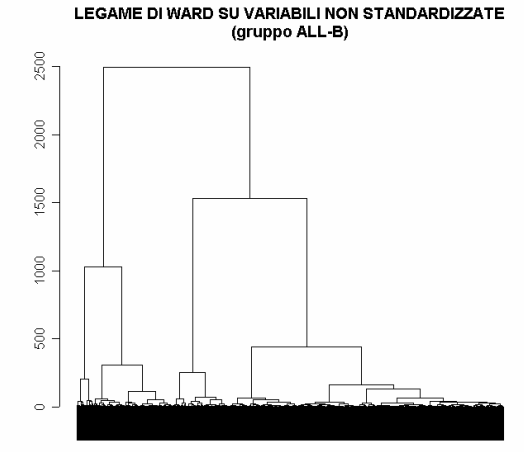

Figura A3.27 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* completo, gruppo ALL-B,variabili non standardizzate.

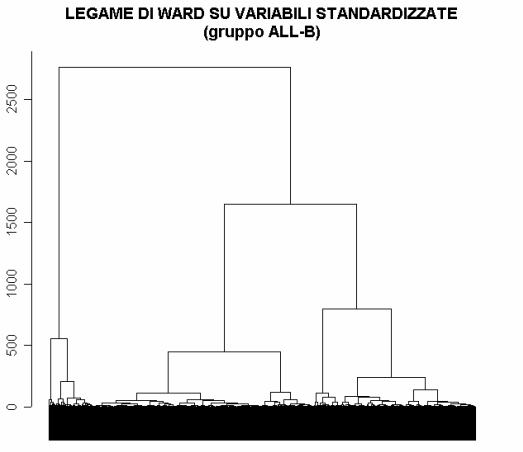

Figura A3.28 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* completo, gruppo ALL-B,variabili standardizzate.

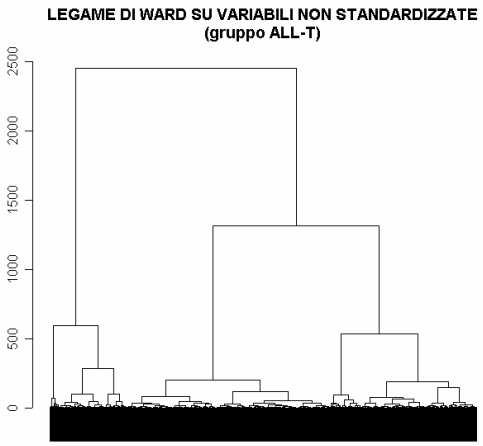

Figura A3.29 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* completo, gruppo ALL-T, variabili non standardizzate.

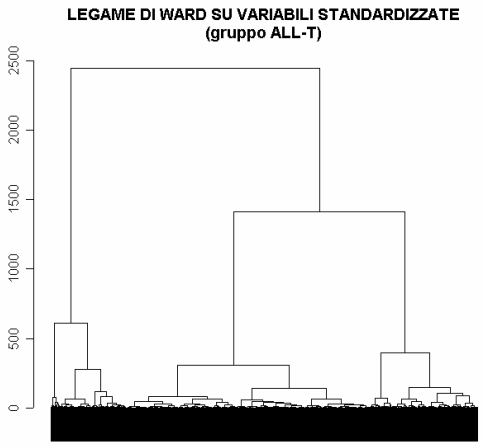

Figura A3.30 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* completo, gruppo ALL-T, variabili standardizzate.

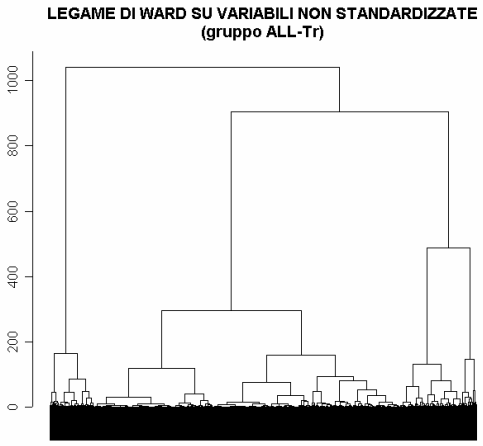

Figura A3.31 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* completo, gruppo ALL-Tr, variabili non standardizzate.

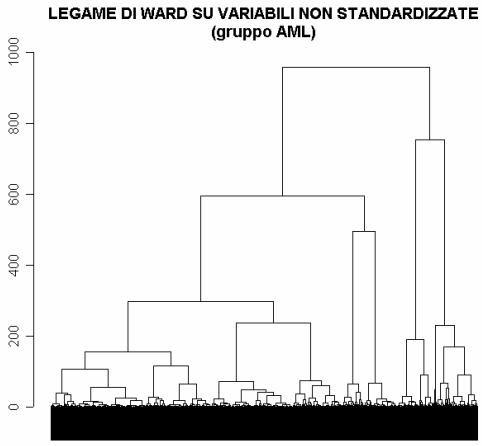

Figura A3.32 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* completo, gruppo AML,variabili non standardizzate.

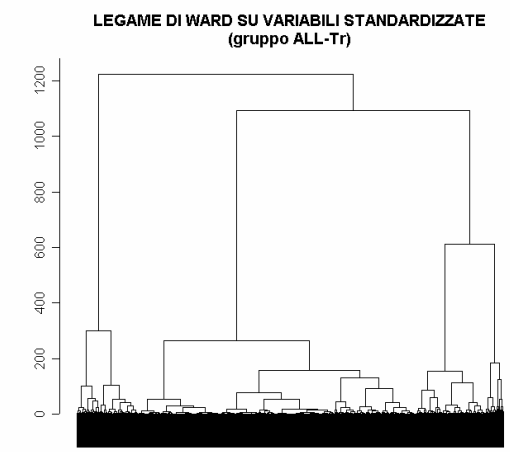

Figura A3.33 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* completo, gruppo ALL-Tr,variabili standardizzate.

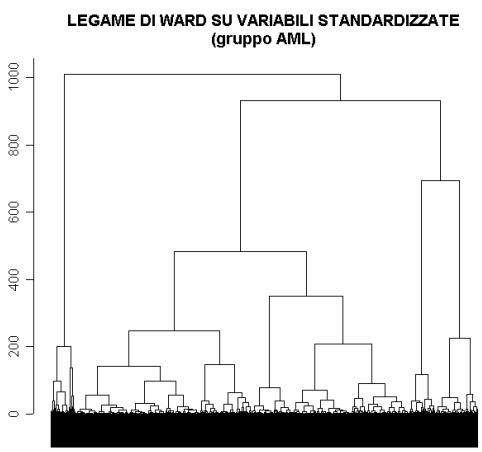

Figura A3.34 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* completo, gruppo AML,variabili standardizzate.

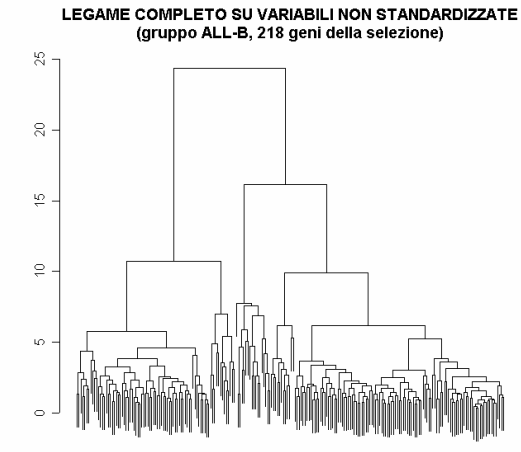

Figura A3.35 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* ridotto, gruppo ALL-B, variabili non standardizzate.

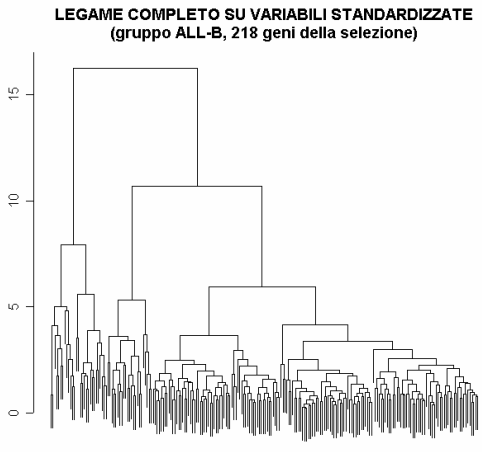

Figura A3.36 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* ridotto, gruppo ALL-B, variabili standardizzate.

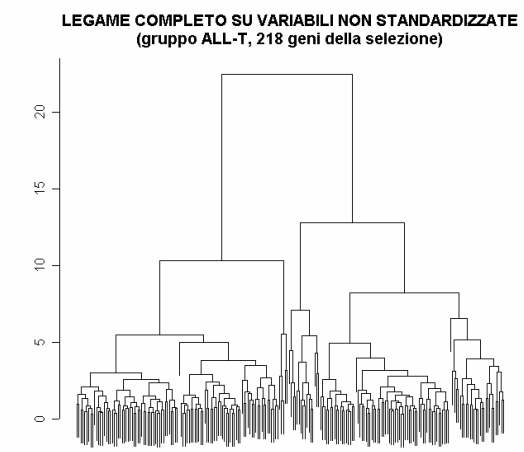

Figura A3.37 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* ridotto, gruppo ALL-T, variabili non standardizzate.

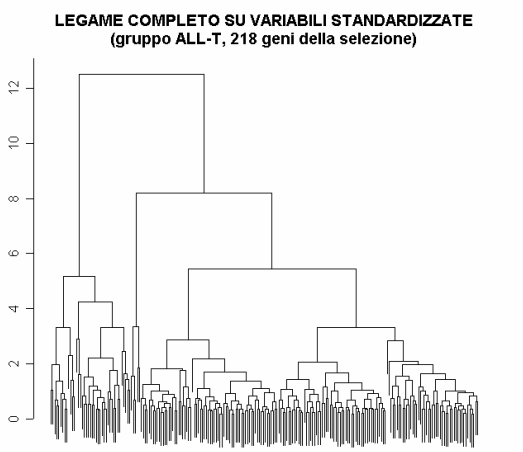

Figura A3.38 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* ridotto, gruppo ALL-T, variabili standardizzate.

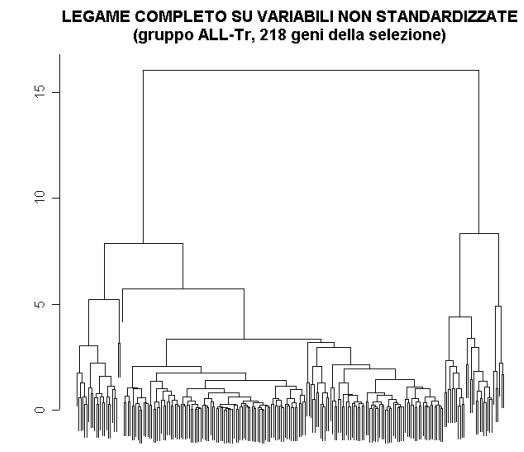

Figura A3.39 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* ridotto, gruppo ALL-Tr, variabili non standardizzate.

Figura A3.40 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* ridotto, gruppo ALL-Tr, variabili standardizzate.

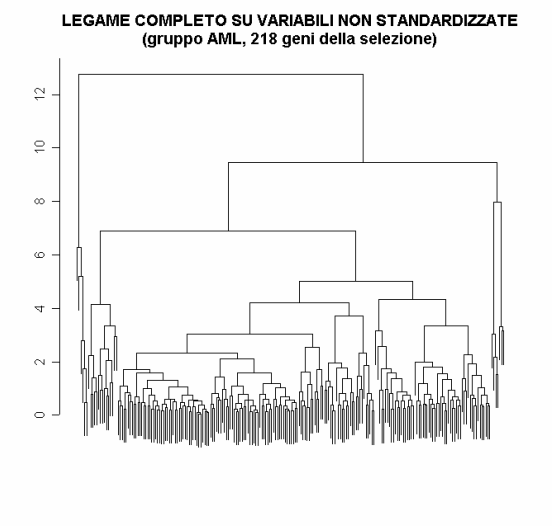

Figura A3.41 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* ridotto, gruppo AML, variabili non standardizzate.

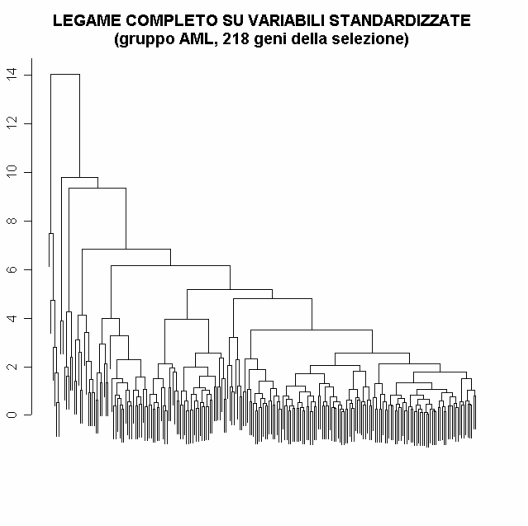

Figura A3.42 Dendrogramma relativo al legame completo. Analisi condotta sul *dataset* ridotto, gruppo AML, variabili standardizzate.

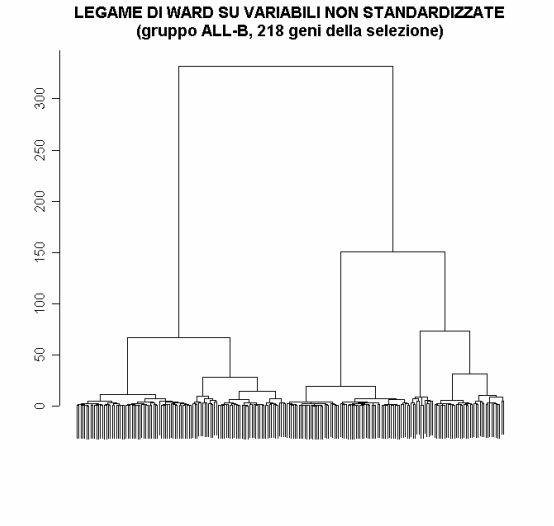

Figura A3.43 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* ridotto, gruppo ALL-B, variabili non standardizzate.

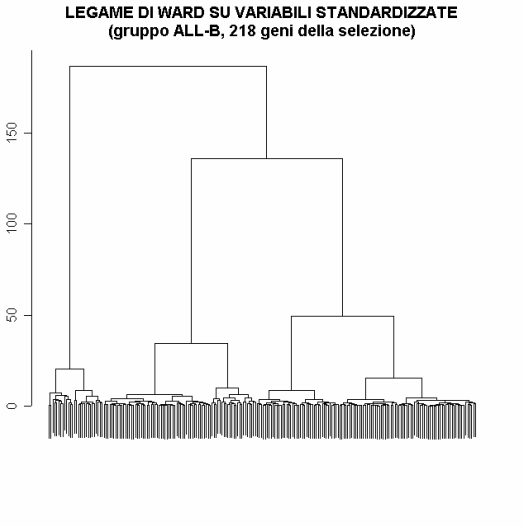

Figura A3.44 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* ridotto, gruppo ALL-B, variabili standardizzate.

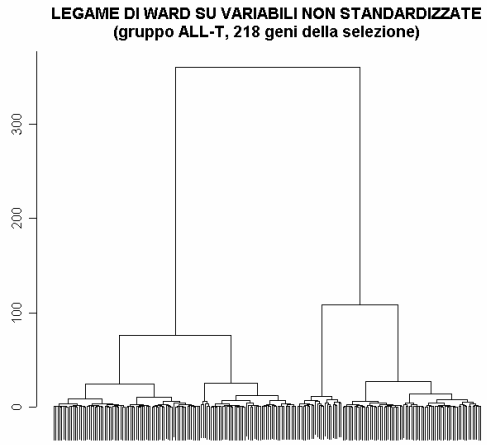

Figura A3.45 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* ridotto, gruppo ALL-T, variabili non standardizzate.

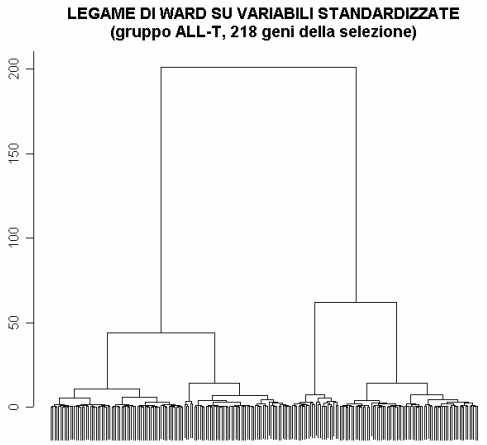

Figura A3.46 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* ridotto, gruppo ALL-T, variabili standardizzate.

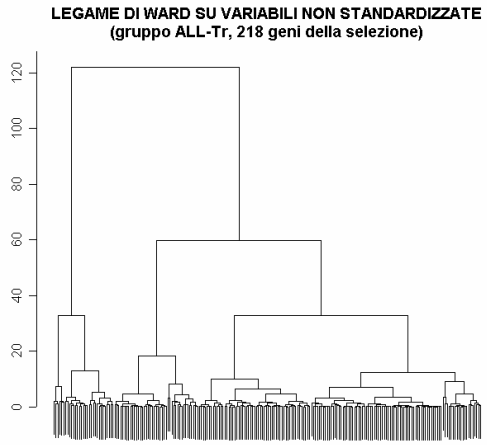

Figura A3.47 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* ridotto, gruppo ALL-Tr, variabili non standardizzate.

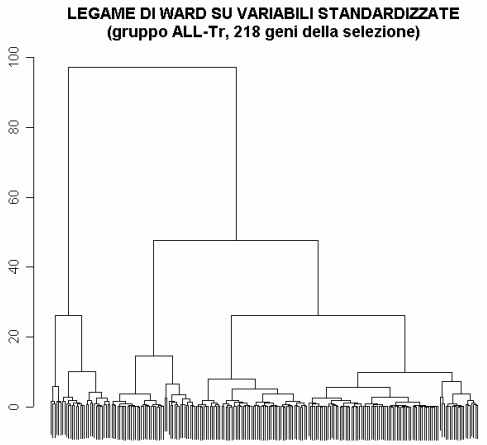

Figura A3.48 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* ridotto, gruppo ALL-Tt, variabili standardizzate.

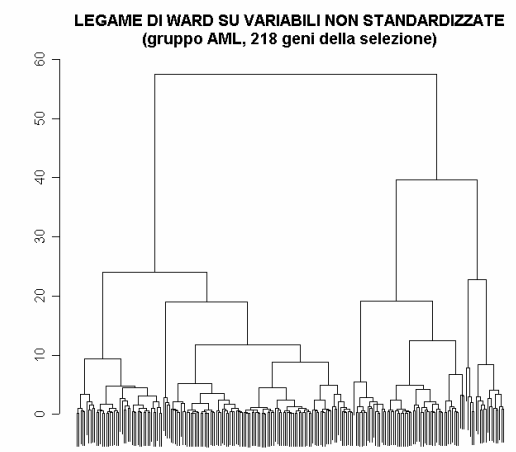

Figura A3.49 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* ridotto, gruppo AML, variabili non standardizzate.

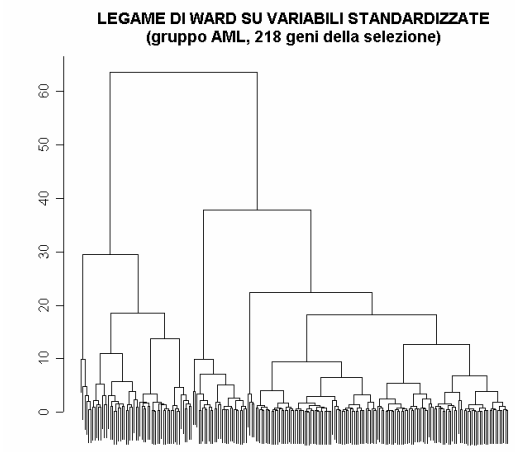

Figura A3.50 Dendrogramma relativo al legame di Ward. Analisi condotta sul *dataset* ridotto, gruppo AML, variabili standardizzate.

A3.4 Confronto tra *dataset* completo e ridotto

Confronto tra i risultati ottenuti applicando le tecniche di *cluster*ing a tutti i geni del *dataset* e la sola selezione (con riferimento alle variabili standardizzate). Ogni gruppo di quattro tabelle si riferisce ad uno dei legami che sono stati utilizzati mentre la singola tabella rappresenta un ciascun gruppo di patologia. Se lette per colonna esse illustrano se e come i geni della selezione che appartengono allo stesso *cluster* sono stati separati quando si sono usati tutti i geni per la classificazione, se lette per riga invece si suggeriscono quanto la classificazione che fa uso di tutti i geni rispecchia quelle che usa solo la selezione.

Metodo k-medie			Classificazione con i geni della selezione			
ALL-B			Cluster			
			1	2	3	4
Classificazione con tutti i geni	Cluster	1	28	0	0	39
		2	0	4	22	0
		3	0	12	0	61
		4	0	52	0	0

(a)

Metodo k-medie			Classificazione con i geni della selezione			
ALL-T			Cluster			
			1	2	3	4
Classificazione con tutti i geni	Cluster	1	0	3	0	46
		2	24	4	0	0
		3	0	65	0	0
		4	0	0	49	27

(b)

Metodo k-medie			Classificazione con i geni della selezione			
ALL-Tr			Cluster			
			1	2	3	4
Classificazione con tutti i geni	Cluster	1	124	4	0	0
		2	0	0	5	6
		3	1	32	0	0
		4	15	0	31	0

(c)

Metodo k-medie			Classificazione con i geni della selezione			
AML			Cluster			
			1	2	3	4
Classificazione con tutti i geni	Cluster	1	106	0	50	0
		2	0	7	13	0
		3	16	0	0	10
		4	16	0	0	0

(d)

Tabella A3.1 (a)-(d) Metodo delle *k*-medie, suddivisione in quattro gruppi: confronto tra l'analisi *cluster* condotta sul *dataset* completo e ridotto.

Legame di Ward		Classificazione con i geni della selezione			
ALL-B		Cluster			
		1	2	3	4
Classificazione con tutti i geni	1	0	64	0	0
Cluster	2	52	14	0	0
	3	18	0	0	41
	4	0	0	28	1

(a)

Legame di Ward		Classificazione con i geni della selezione			
ALL-T		Cluster			
		1	2	3	4
Classificazione con tutti i geni	1	42	1	55	0
Cluster	2	25	0	0	0
	3	1	69	0	1
	4	0	1	0	23

(b)

Legame di Ward		Classificazione con i geni della selezione			
ALL-Tr		Cluster			
		1	2	3	4
Classificazione con tutti i geni	1	3	33	0	0
Cluster	2	104	10	0	0
	3	3	0	6	18
	4	34	0	0	7

(c)

Legame di Ward		Classificazione con i geni della selezione			
AML		Cluster			
		1	2	3	4
Classificazione con tutti i geni	1	9	8	0	0
Cluster	2	121	8	11	0
	3	3	35	1	6
	4	0	0	16	0

(d)

Tabella A3.2 (a)-(d) Legame di Ward, suddivisione in quattro gruppi: confronto tra l'analisi *cluster* condotta sul *dataset* completo e ridotto.

Metodo dei medoidi		Classificazione con i geni della selezione			
ALL-B		Cluster			
		1	2	3	4
Classificazione con tutti i geni	1	4	74	0	0
Cluster	2	45	1	0	0
	3	27	0	29	0
	4	0	0	19	19

(a)

Metodo dei medoidi		Classificazione con i geni della selezione			
ALL-T		Cluster			
		1	2	3	4
Classificazione con tutti i geni	1	9	0	72	0
Cluster	2	30	0	0	0
	3	0	52	0	27
	4	21	7	0	0

(b)

Metodo dei medoidi		Classificazione con i geni della selezione			
ALL-Tr		Cluster			
		1	2	3	4
Classificazione con tutti i geni	1	2	39	0	0
Cluster	2	99	7	0	0
	3	30	0	0	15
	4	0	0	6	20

(c)

Metodo dei medoidi		Classificazione con i geni della selezione			
AML		Cluster			
		1	2	3	4
Classificazione con tutti i geni	1	24	3	7	0
Cluster	2	39	2	76	1
	3	0	38	4	1
	4	0	0	0	23

(d)

Tabella A3.3 (a)-(d) Metodo dei medoidi, suddivisione in quattro gruppi: confronto tra l'analisi *cluster* condotta sul *dataset* completo e ridotto

Legame di Ward		Classificazione con i geni della selezione		
ALL-B		Cluster		
		1	2	3
Classificazione con tutti i geni: Cluster	1	0	64	0
	2	52	14	0
	3	60	0	28

(a)

Legame di Ward		Classificazione con i geni della selezione		
ALL-T		Cluster		
		1	2	3
Classificazione con tutti i geni: Cluster	1	97	1	0
	2	25	0	0
	3	1	70	24

(b)

Legame di Ward		Classificazione con i geni della selezione		
ALL-Tr		Cluster		
		1	2	3
Classificazione con tutti i geni: Cluster	1	3	33	0
	2	104	10	0
	3	37	0	31

(c)

Legame di Ward		Classificazione con i geni della selezione		
AML		Cluster		
		1	2	3
Classificazione con tutti i geni: Cluster	1	130	16	11
	2	3	41	1
	3	0	0	16

(d)

Tabella A3.4 (a)-(d) Legame di Ward, suddivisione in tre gruppi: confronto tra l'analisi *cluster* condotta sul *dataset* completo e ridotto.

Metodo k-medie		Classificazione con i geni della selezione		
ALL-B		Cluster		
		1	2	3
Classificazione con tutti i geni: Cluster	1	79	0	0
	2	0	24	21
	3	23	0	71

(a)

Metodo k-medie		Classificazione con i geni della selezione		
ALL-T		Cluster		
		1	2	3
Classificazione con tutti i geni: Cluster	1	27	43	0
	2	0	0	80
	3	0	42	26

(b)

Metodo k-medie		Classificazione con i geni della selezione		
ALL-Tr		Cluster		
		1	2	3
Classificazione con tutti i geni: Cluster	1	15	0	134
	2	25	6	0
	3	0	0	38

(c)

Metodo k-medie		Classificazione con i geni della selezione		
AML		Cluster		
		1	2	3
Classificazione con tutti i geni: Cluster	1	1	21	0
	2	16	0	7
	3	47	126	0

(d)

Tabella A3.5 (a)-(d) Metodo dell *k*-medie, suddivisione i tre gruppi: confronto tra l'analisi *cluster* condotta sul *dataset* completo e ridotto.

Legame completo		Classificazione con i geni della selezione		
ALL-B		Cluster		
		1	2	3
Classificazione con tutti i geni	Cluster 1	101	0	23
	Cluster 2	65	14	0
	Cluster 3	0	15	0

(a)

Legame completo		Classificazione con i geni della selezione		
ALL-T		Cluster		
		1	2	3
Classificazione con tutti i geni	Cluster 1	23	0	5
	Cluster 2	149	18	0
	Cluster 3	0	23	0

(b)

Legame completo		Classificazione con i geni della selezione		
ALL-Tr		Cluster		
		1	2	3
Classificazione con tutti i geni	Cluster 1	69	0	0
	Cluster 2	118	0	24
	Cluster 3	0	6	1

(c)

Legame completo		Classificazione con i geni della selezione		
AML		Cluster		
		1	2	3
Classificazione con tutti i geni	Cluster 1	145	0	0
	Cluster 2	65	1	2
	Cluster 3	0	5	0

(d)

Tabella A3.6 (a)-(d) Legame completo, suddivisione i tre gruppi: confronto tra l'analisi *cluster* condotta sul *dataset* completo e ridotto.

Metodo dei medoidi		Classificazione con i geni della selezione		
ALL-B		Cluster		
		1	2	3
Classificazione con tutti i geni	Cluster 1	0	79	0
	Cluster 2	72	6	0
	Cluster 3	33	0	28

(a)

Metodo dei medoidi		Classificazione con i geni della selezione		
ALL-T		Cluster		
		1	2	3
Classificazione con tutti i geni	Cluster 1	12	0	72
	Cluster 2	50	0	0
	Cluster 3	6	78	0

b

Metodo dei medoidi		Classificazione con i geni della selezione		
ALL-Tr		Cluster		
		1	2	3
Classificazione con tutti i geni	Cluster 1	2	39	0
	Cluster 2	112	7	0
	Cluster 3	17	0	41

(c)

Metodo dei medoidi		Classificazione con i geni della selezione		
AML		Cluster		
		1	2	3
Classificazione con tutti i geni	Cluster 1	60	9	0
	Cluster 2	88	36	6
	Cluster 3	0	0	19

(d)

Tabella A3.7 (a)-(d) Metodo dei medoidi, suddivisione i tre gruppi: confronto tra l'analisi *cluster* condotta sul *dataset* completo e ridotto.

A3.5 Curve di Andrews: liste di geni

Tabelle riguardantile liste dei geni che fuoriescono dalle bande fissate a varie percentuali (Tabella A3.8 e Tabella A3.9) e a diversi livelli (Tabella A3.10 e Tabella A3.11). Le Tabella A3.8 e A.3.10 si riferiscono ai geni che generano curve di Andrews che fuoriescono da tali bande per un solo gruppo, Tabella A3.9 e A.3.11 ai geni che generano curve di Andrews che *non* fuoriescono dalle bande per un solo gruppo.

Geni le cui curve di Andrews fuoriescono dalle bande per una sola classe																
Livello 1%																
ALL-B	108	130	539	598	898	905	1857	1977	2481	3127	3437	4076	4167	4190	4262	
	4525	4535	4673	4831												
ALL-T	990	1402	1505	1513	1691	2454	2735	2765	2868	3129	3156	3224	3489	3828	4215	
	4380	4463	4478	4489	4697	4840										
ALL-Tr	10	30	70	97	149	226	1342	1645	1673	1688	1907	2720	2821	3298	3852	
	4077	4110	4174	4189	4854											
AML	65	349	482	524	628	1048	1560	1736	2032	2148	2235	2477	2715	3000	3056	
	3581	3721	3850	4028	4080											
Livello 3%																
ALL-B	100	108	111	114	128	228	259	539	561	562	582	742	853	891	898	
	969	1125	1180	1225	1437	1592	1643	2277	2281	2481	2553	3312	3725	3953	4060	
	4101	4128	4163	4164	4433	4443	4534	4535	4673	4699						
ALL-T	202	204	223	235	239	286	409	420	487	560	902	1107	1222	1347	1448	
	1636	1646	1816	2134	2454	2616	2695	2710	2748	2812	2868	2869	3001	3088	3156	
	3310	3401	3449	4142	4162	4415	4478	4480	4489	4653	4687	4697	4837	4840		
ALL-Tr	30	32	33	67	71	73	93	96	106	113	122	125	142	206	221	
	257	258	401	405	411	491	502	507	565	601	906	936	1001	1066	1085	
	1168	1230	1265	1351	1576	1607	1695	1709	1718	1723	1874	1938	1940	2174	2206	
	2466	3210	3294	3298	3483	3801	3841	3862	3869	3873	3954	4039	4067	4089	4111	
	4334	4441	4470	4643	4669	4677	4826									
AML	31	76	123	138	230	349	435	461	482	524	527	628	633	640	648	
	985	1065	1071	1112	1142	1149	1206	1328	1427	1541	1588	1683	1687	1707	2032	
	2148	2217	2228	2229	2355	2399	2477	2715	2771	2802	2977	3000	3084	3144	3255	
	3345	3374	3721	3773	3971	4093	4115	4186	4227	4235	4249	4325	4465	4482	4585	
	4710	4723	4762	4912	4915											

Livello 5%															
ALL-B	145	147	171	279	694	853	891	898	981	1125	1131	1134	1343	1413	1445
	1570	2019	2161	2209	2481	2511	2714	2805	2813	3106	3725	3731	3965	4002	4040
	4060	4137	4141	4164	4166	4172	4337	4357	4369	4433	4483	4507	4534	4535	4612
	4627	4638	4673	4753	4822	4868									
ALL-T	392	400	402	486	487	499	513	535	599	665	839	895	896	952	1033
	1244	1347	1355	1399	1409	1429	1626	1712	1816	1909	2030	2354	2465	2616	2637
	2673	2695	2697	2738	2748	2804	2869	3088	3126	3156	3203	3346	3418	3425	3442
	3454	3601	3806	3855	4015	4043	4142	4200	4201	4326	4364	4393	4478	4480	4481
	4489	4660	4754	4756	4837	4840									
ALL-Tr	18	25	39	60	67	71	73	92	93	106	142	159	196	209	217
	221	282	308	314	344	346	491	492	498	501	507	518	526	545	565
	601	659	669	910	913	931	936	939	950	951	1106	1113	1155	1230	1239
	1378	1432	1449	1511	1553	1579	1607	1664	1669	1671	1695	1699	1706	1723	1730
	1874	1897	1901	1923	1938	2042	2053	2081	2146	2174	2202	2205	2206	2482	2649
	2726	2982	3059	3206	3257	3296	3483	3570	3588	3671	3841	3844	3847	3862	3864
	3867	3869	3873	3876	3886	3954	3962	4001	4042	4059	4089	4114	4156	4161	4184
	4196	4229	4254	4470	4474	4502	4547	4677							
AML	76	79	84	123	138	160	256	304	349	356	389	425	461	482	522
	524	527	538	569	613	617	628	633	640	648	928	985	1017	1065	1071
	1112	1142	1146	1149	1231	1328	1472	1541	1547	1588	1620	1621	1687	1707	1976
	2032	2148	2182	2217	2228	2229	2355	2398	2399	2477	2542	2771	2779	2802	2907
	2977	3000	3062	3082	3084	3092	3144	3195	3255	3319	3345	3374	3410	3553	3694
	3721	3745	3773	3818	3842	3866	3971	4035	4115	4157	4186	4187	4191	4211	4233
	4240	4249	4250	4252	4270	4352	4367	4374	4465	4482	4493	4585	4710	4723	4762
	4787	4839	4880	4912	4915										
Livello 10%															
ALL-B	26	28	145	183	459	531	664	694	804	886	891	912	921	923	942
	946	1098	1125	1226	1426	1433	1451	1459	1570	1708	1710	2199	2288	2304	2433
	2481	2681	2714	2805	2810	2848	2953	2988	2991	3043	3106	3107	3143	3241	3303
	3342	3364	3377	3390	3531	3623	3725	3737	3791	3809	3872	3999	4026	4048	4060
	4066	4096	4119	4214	4222	4265	4341	4356	4396	4407	4412	4460	4472	4512	4535
	4555	4603	4638	4655	4694	4714	4735	4760	4785	4798	4816	4822			
ALL-T	49	207	487	505	535	570	611	668	671	720	729	771	793	817	852
	884	885	1033	1144	1150	1185	1223	1310	1316	1333	1374	1409	1429	1482	1567
	1580	1582	1637	1638	1650	1728	1816	1835	1858	1912	2017	2244	2268	2302	2351
	2423	2438	2455	2468	2616	2697	2804	2849	2916	2940	3042	3073	3080	3103	3110
	3156	3253	3321	3418	3555	3601	3621	3629	3710	3958	4011	4021	4043	4103	4142
	4178	4200	4275	4324	4438	4447	4480	4518	4520	4625	4648	4701	4764	4768	4780
	4808	4840	4863	4867											
	1611	1623	1634	1676	1693	1702	1703	1706	1730	1741	1833	1878	1941	1958	2042
	2046	2047	2053	2106	2145	2166	2169	2201	2245	2262	2294	2317	2382	2390	2403
	2622	2649	2749	2763	2910	2962	3091	3212	3257	3296	3341	3385	3393	3438	3483
	3570	3647	3703	3841	3843	3844	3845	3846	3849	3851	3857	3862	3863	3867	3870
	3873	3876	3877	3880	3886	3938	3962	3969	3987	3993	3998	4016	4023	4041	4059
	4063	4072	4085	4086	4112	4125	4156	4161	4184	4203	4209	4210	4255	4259	4267
	4276	4290	4323	4331	4395	4409	4452	4502	4530	4547	4667	4677	4761	4777	4779
AML	64	76	77	123	175	178	190	304	310	339	349	356	359	386	425
	461	481	482	522	524	527	536	537	538	613	617	633	635	640	648
	1875	1885	1905	1959	1976	2003	2010	2095	2108	2148	2219	2228	2229	2343	2355
	2360	2389	2398	2399	2402	2432	2477	2535	2552	2560	2561	2565	2640	2668	2677
	3467	3539	3665	3694	3721	3745	3772	3773	3792	3793	3803	3818	3932	3971	3975
	3982	4027	4035	4046	4058	4074	4088	4113	4115	4118	4146	4157	4191	4202	4233

Tabella A3.8 Geni che generano profili di Andrews che fuoriescono per un solo gruppo dalle bande fissate a varie percentuali.

Geni le cui curve di Andrews non fuoriescono dalle bande per una sola classe																
Livello 1%																
ALL-B	566	4033														
ALL-T	36	43	153	433	1213	1545	2015	2704	2949	4271						
ALL-Tr	225	1533	2743	2866	3045	3624	3995	4149	4416	4713						
AML	452	1314	4018	4030	4102	4107	4133	4244	4859							
Livello 3%																
ALL-B	1505	1513														
ALL-T	70	97	121	149	511	1515	1545	1560	4226	4645	4854					
ALL-Tr	161	210	262	439	563	932	1109	1300	1533	1691	2743	2794	3489	3619	4176	
	4340	4428														
AML	130	226	397	449	452	558	990	1673	1688	1721	2984	3127	3286	3810	3967	
	3996	4077	4100	4102	4105	4107	4110	4215	4445	4521	4525	4831				
Livello 5%																
ALL-B	202	458	3310													
ALL-T	32	33	70	97	121	149	1066	1515	1625	1647	1931	1943	2171	2524	2715	
	2720	3852	3961	4039	4153	4159	4174	4334	4696							
ALL-Tr	102	210	222	223	228	230	235	239	263	439	562	900	1012	1107	1300	
	1636	1646	1691	2230	2277	2765	2794	3129	3489	3556	3619	4258	4881			
AML	89	108	113	195	205	286	558	902	938	990	1110	1168	1265	1402	1473	
	1592	1688	1721	1940	2129	2141	2237	2281	2466	2769	2984	3283	3308	3312	3449	
	3527	3801	3810	4003	4067	4077	4100	4101	4105	4107	4110	4111	4167	4266	4445	
	4521	4669	4794	4818	4886											
Livello 10%																
ALL-B	81	144	196	198	445	486	931	1155	1427	1621	1683	1933	2013	2346	2542	
	2726	2735	3162	4321	4441	4880										
ALL-T	48	60	74	125	168	173	282	308	411	431	582	819	1342	1351	1423	
	1455	1722	1723	1931	2202	3056	3262	3294	3436	3954	4185	4189	4252	4646	4872	
ALL-Tr	1282	1328	1347	1355	1399	1412	1620	1915	1951	2181	2230	2277	2467	2695	2771	
	2814	2868	3129	3203	3224	3316	3499	3828	3953	3955	3989	4013	4032	4073	4140	
AML	27	45	92	140	147	151	191	205	400	420	436	498	499	513	539	
	565	637	670	742	839	898	902	906	927	937	950	970	1110	1169	1222	
	4386	4411	4433	4473	4509	4534	4627	4653	4697	4738	4781	4794	4812	4818	4858	
Livello 15%																
ALL-B	25	29	122	386	463	553	556	578	826	1079	1106	1139	1155	1187	1427	
	1618	1837	1878	1996	2013	2235	2346	2542	2865	3162	3415	3824	3858	4041	4765	
ALL-T	37	84	88	124	132	159	173	179	221	282	399	459	471	501	507	
	582	630	985	1082	1342	1445	1455	1456	1718	1890	1931	2042	2146	2207	2482	
	3056	3262	3279	3436	3646	3866	3939	4082	4116	4186	4259	4346	4405	4649	4872	
	4889															
ALL-Tr	40	49	98	111	204	212	216	228	231	235	239	242	256	259	263	
	280	287	393	432	531	576	628	668	800	867	880	885	900	915	946	
	952	1065	1098	1146	1148	1178	1206	1272	1282	1433	1617	1620	1887	1892	1927	
	1951	2180	2181	2277	2416	2423	2426	2467	2492	2771	2814	2868	3117	3129	3203	
	3224	3241	3322	3499	3685	3982	4002	4017	4032	4036	4043	4073	4140	4162	4180	
	4188	4201	4211	4233	4328	4415	4429	4432	4442	4463	4493	4507	4518	4520	4568	
	4592	4617	4731	4868												
AML	45	57	92	151	191	209	224	294	357	416	456	498	539	554	565	
	637	643	653	757	830	906	923	937	949	950	970	1025	1113	1125	1222	
	1244	1309	1310	1333	1413	1440	1543	1557	1561	1643	1923	1924	1945	1986	2016	
	2019	2027	2029	2030	2043	2071	2119	2137	2161	2206	2263	2304	2409	2465	2466	
	2486	2517	2553	2673	2710	2722	2740	2748	2812	2848	2957	2958	3001	3103	3104	
	3192	3206	3207	3268	3312	3342	3346	3442	3454	3483	3570	3588	3801	3861	3864	
	3899	3963	3974	4014	4015	4037	4045	4090	4091	4111	4130	4136	4137	4141	4144	
	4181	4196	4217	4229	4342	4357	4369	4409	4411	4433	4443	4470	4509	4534	4537	
	4622	4697	4717	4735	4738	4769	4785	4812	4818	4855	4858	4874				

Tabella A3.9 Geni che generano profili di Andrews che non fuoriescono per un solo gruppo dalle bande fissate a varie percentuali.

Geni le cui curve di Andrews fuoriescono dalle bande per una sola classe	
Livello ± 30	
ALL-B	1213 1314 2690 3045 3125 3937 4245 4247 4416
ALL-T	
ALL-Tr	
AML	
Livello ± 28	
ALL-B	1213 1271 1303 1314 2690 2866 3045 3125 3128 3937 4029 4030 4149 4245 4247 4416 4589 4700
ALL-T	
ALL-Tr	
AML	
Livello ± 26	
ALL-B	225 452 454 1213 1271 1303 1314 2690 2866 2927 2949 3045 3125 3128 3624 3937 3996 4029 4030 4102 4149 4245 4247 4271 4416 4589 4700 4713
ALL-T	
ALL-Tr	
AML	
Livello ± 24	
ALL-B	225 433 452 454 1213 1271 1276 1303 1314 2690 2866 2927 2949 3045 3125 3128 3466 3624 3937 3996 4018 4029 4030 4038 4097 4100 4102 4107 4108 4133 4149 4167 4168 4245 4247 4271 4416 4589 4700 4713
ALL-T	
ALL-Tr	
AML	
Livello ± 22	
ALL-B	225 433 452 454 455 876 905 1213 1271 1276 1303 1314 2690 2743 2866 2927 2949 3045 3125 3128 3437 3466 3624 3775 3802 3937 3996 4018 4029 4030 4038 4049 4076 4097 4100 4102 4107 4108 4133 4149 4167 4168 4190 4245 4247 4262 4271 4416 4525 4589 4700 4713
ALL-T	
ALL-Tr	
AML	
Livello ± 20	
ALL-B	225 433 451 452 454 455 876 900 905 916 1213 1271 1276 1303 1314 1365 2690 2743 2866 2927 2949 3045 3125 3128 3404 3437 3466 3489 3619 3624 3775 3802 3937 3996 4009 4018 4029 4030 4038 4049 4076 4097 4100 4102 4107 4108 4110 4133 4149 4167 4168 4190 4198 4215 4245 4247 4262 4271 4360 4416 4428 4445 4525 4589 4633 4700 4713
ALL-T	
ALL-Tr	
AML	
Livello ± 18	
ALL-B	10 59 116 225 417 433 451 452 454 455 525 539 876 900 905 916 990 1213 1271 1276 1300 1303 1365 1645 2690 2743 2866 2927 2949 2965 3125 3128 3139 3224 3404 3437 3466 3489 3535 3581 3619 3624 3775 3802 3937 3996 4009 4018 4028 4029 4030 4038 4049 4076 4080 4097 4100 4102 4105 4107 4108 4110 4133 4167 4168 4190 4198 4215 4223 4244 4245 4247 4262 4271 4360 4416 4428 4445 4508 4525 4589 4633 4700 4869
ALL-T	
ALL-Tr	
AML	

Livello ± 16															
ALL-B	10	59	116	161	225	226	417	433	451	452	454	455	525	539	557
	558	559	561	566	876	879	900	905	916	990	1213	1271	1276	1300	1365
	1645	2690	2927	2949	2965	2984	3127	3128	3129	3139	3224	3283	3334	3404	3437
	3466	3489	3535	3581	3619	3624	3775	3802	3937	3945	4009	4018	4028	4029	4033
	4038	4049	4076	4077	4080	4097	4100	4105	4106	4107	4108	4110	4133	4167	4168
	4190	4198	4215	4223	4244	4245	4262	4271	4360	4428	4445	4473	4508	4525	4535
	4633	4645	4700	4869											
ALL-T															
ALL-Tr															
AML															
Livello ± 14															
ALL-B	10	36	40	59	116	130	153	161	172	203	210	226	228	262	397
	398	417	431	433	439	451	453	455	511	525	539	557	558	559	561
	562	563	566	658	797	876	879	898	900	905	990	1048	1213	1252	1271
	1282	1300	1365	1385	1473	1533	1545	1560	1645	1666	1673	1686	1736	1930	2015
	2230	2481	2689	2690	2704	2739	2794	2949	2965	2984	3127	3129	3139	3283	3334
	3404	3437	3466	3489	3499	3535	3556	3581	3619	3775	3810	3945	3967	3995	4009
	4018	4028	4029	4033	4049	4076	4077	4080	4100	4105	4106	4108	4110	4167	4168
	4177	4190	4198	4215	4223	4226	4244	4262	4266	4271	4340	4360	4428	4445	4473
	4508	4525	4535	4633	4645	4646	4673	4699	4859	4869					
ALL-T	3828														
ALL-Tr															
AML															
Livello ± 12															
ALL-B	10	36	40	43	59	108	116	128	130	153	161	164	172	203	204
	205	210	222	223	226	228	259	262	263	397	398	417	420	431	433
	439	449	453	511	525	539	557	558	559	561	562	563	598	658	742
	797	819	867	876	879	898	900	905	915	969	1012	1048	1068	1109	1218
	1252	1265	1282	1300	1365	1385	1402	1473	1515	1533	1545	1560	1594	1614	1625
	1643	1645	1666	1673	1677	1686	1736	1857	1930	1977	2015	2230	2281	2481	2689
	2704	2739	2769	2794	2821	2949	2965	2984	3127	3283	3286	3312	3326	3334	3404
	3437	3499	3535	3556	3581	3619	3683	3775	3945	3953	3967	3995	4003	4009	4028
	4032	4033	4076	4077	4080	4104	4106	4108	4110	4163	4164	4167	4177	4180	4190
	4221	4223	4226	4241	4244	4262	4266	4271	4340	4432	4433	4443	4445	4463	4473
	4508	4525	4535	4601	4645	4646	4673	4687	4699	4818	4820	4831	4859	4868	4869
ALL-T	3828	4489													
ALL-Tr															
AML															

Tabella A3.10 Geni che generano profili di Andrews che fuoriescono per un solo gruppo dalle bande fissate a vari livelli

Geni le cui curve di Andrews fuoriescono dalle bande per una sola classe	
Livello ± 14	
ALL-B	
ALL-T	
ALL-Tr	454 3937 4416 4700
AML	
Livello ± 12	
ALL-B	
ALL-T	
ALL-Tr	454 1303 3128 3937 4360 4416 4700
AML	4107
Livello ± 10	
ALL-B	
ALL-T	4009
ALL-Tr	433 455 916 2690 3045 3125 3128 3466 3802 3937 4360 4416 4700
AML	417 1314 2949 3775 4018 4105 4107 4589
Livello ± 8	
ALL-B	
ALL-T	
ALL-Tr	433 1365 1736 3045 3466 4028 4080 4106 4108 4360 4416 4633 4713
AML	10 452 1314 3535 3775 4018 4030 4077 4105 4107 4110 4133 4589
Livello ± 6	
ALL-B	
ALL-T	4189
ALL-Tr	116 153 172 225 262 435 439 525 1048 1271 1533 1560 1736 2704 2743 2866 3045 3581 3624 3828 3995 4028 4029 4076 4080 4097 4108 4149 4177 4428 4633
AML	226 452 511 1673 1677 3996 4030 4077 4102 4105 4107 4110 4133
Livello ± 4	
ALL-B	
ALL-T	35 65 125 1066 2171 3056 4189 4741 4872
ALL-Tr	31 79 102 111 114 158 161 165 210 228 230 239 262 263 397 435 439 449 458 557 559 562 563 628 900 932 997 1012 1048 1109 1149 1206 1218 1252 1300 1328 1385 1513 1533 1625 1646 1657 1691 1930 2023 2181 2557 2725 2735 2743 2765 2771 2794 2866 2976 3129 3286 3311 3437 3489 3556 3563 3581 3619 3828 3961 3967 4149 4159 4176 4177 4221 4223 4227 4235 4258 4325 4334 4340 4380 4428 4745 4826 4831 4869 4880
AML	89 106 195 205 206 452 565 1614 1688 1709 1940 2206 2720 2984 3214 3298 3801 3852 4003 4067 4077 4104 4105 4107 4110 4130 4215 4445 4521 4801 4886

Tabella A3.11 Geni che generano profili di Andrews che non fuoriescono per un solo gruppo dalle bande fissate a vari livelli.

A3.6 *Routines* R

A3.5.r1 `#confronta i geni a seconda che le rispettive curve di Andrews rimangono all'interno o escano dalle bande`

```
creamat<-function (A,B,C,D)

{
x<-0
x1<-NULL
x2<-NULL
x3<-NULL
a1<-NULL
b1<-NULL
c1<-NULL
d1<-NULL
a2<-NULL
b2<-NULL
c2<-NULL
d2<-NULL
va<-rep(0,4992)
vb<-rep(0,4992)
vc<-rep(0,4992)
vd<-rep(0,4992)
va[A]<-1
vb[B]<-1
vc[C]<-1
vd[D]<-1
s<-va+vb+vc+vd
j<-0
k1<-0
k2<-0
k3<-0
for (i in 1:length(va)){
    j<-j+1
    if(s[i]==1){
      k1<-k1+1
      x1[k1]<-j}
    if(s[i]==3){
      k3<-k3+1
      x3[k3]<-j}
    if(s[i]==2){
      k2<-k2+1
      x2[k2]<-j}
   }
j<-0
j1<-0
j2<-0
j3<-0
j4<-0
m<-0
if(length(x1)>0){
for (i in 1:length(x1)){
    j<-j+1
```

```
    m<-x1[i]
    if(va[m]==1){
      j1<-j1+1
      a1[j1]<-m}
    if(vb[m]==1){
      j2<-j2+1
      b1[j2]<-m}
    if(vc[m]==1){
      j3<-j3+1
      c1[j3]<-m}
    if(vd[m]==1){
      j4<-j4+1
      d1[j4]<-m}
   }
}

j<-0
j1<-0
j2<-0
j3<-0
j4<-0
if(length(x3)>0){
for (i in 1:length(x3)){
    j<-j+1
    m<-x3[i]
    if(va[m]==0){
      j1<-j1+1
      a2[j1]<-m}
    if(vb[m]==0){
      j2<-j2+1
      b2[j2]<-m}
    if(vc[m]==0){
      j3<-j3+1
      c2[j3]<-m}
    if(vd[m]==0){
      j4<-j4+1
      d2[j4]<-m}
   }
}

list(ALLB1=a1,ALLT1=b1,ALLTr1=c1,AML1=d1,ALLB3=a2,ALLT3=
b2,ALLTr3=c2,AML3=d
2)

}
```

A3.5.r2

```
#Grafici con pallini

out.plot<-function (b,col=1)
{
a<-as.numeric(b)
x<-0
y<-0
for (i in 1:length(a)){
    x[i]<-ceiling(a[i]/78)
```

```
    y[i]<-a[i]-((ceiling(a[i]/78)-1)*78)
}
plot(x,y,col=col,pch=19,xlab="",ylab="",cex=1,xlim=c(1,6
4),ylim=c(1,78))
}
```

A3.5.r3

```
#Grafici con pallini

out.points<-function (b,col=1,pch=21)
{
a<-as.numeric(b)
x<-0
y<-0
for (i in 1:length(a)){
    x[i]<-ceiling(a[i]/78)
    y[i]<-a[i]-((ceiling(a[i]/78)-1)*78)
}
points(x,y,col=col,pch=pch,cex=1)
}
```

A3.5.r4

```
#grafici con pallini per il dataset ridotto

out.plot200<-function (b,col=1)
{
a<-as.numeric(b)
x<-0
y<-0
for (i in 1:length(a)){
    x[i]<-ceiling(a[i]/11)
    y[i]<-a[i]-((ceiling(a[i]/11)-1)*11)
}
plot(x,y,col=col,pch=19,xlab="",ylab="",cex=1)
}
#,xlim=c(1,11),ylim=c(1,20)
```

A3.5.r5

```
out.points200<-function (b,col=1,pch=21)

{
a<-as.numeric(b)
x<-0
y<-0
for (i in 1:length(a)){
    x[i]<-ceiling(a[i]/11)
    y[i]<-a[i]-((ceiling(a[i]/11)-1)*11)
}
points(x,y,col=col,pch=pch,cex=1)
}
```

A3.5.r6

```
#Standardizzazione delle variabili

stand<-function(dati){
```

```
media<-apply(dati,2,mean)
v<-(var(dati))
sqm<-sqrt(diag(v))
dati.st<-dati
for(i in 1:ncol(dati)){
   dati.st[,i]<-(dati[,i]-media[i])/sqm[i]
   }
dati.st
}
```

A3.5.r7 `#Grafici sul raggruppamento dei geni in` ***`cluster`*** `(4 gruppi)`

```
clusterpoint<-function
(a,b,c,d,cl,gruppo,c1,c2,c3,c4,ngen)
{
if(gruppo==2){
  x0<-5
  x1<-10}
if(gruppo==3){
  x0<-10
  x1<-15}
if(gruppo==4){
  x0<-15
  x1<-20}
x<-0
iniz<-NULL
fine<-NULL
col<-NULL
cola<-NULL
colb<-NULL
colc<-NULL
cold<-NULL
col[cl==1]<-c1
col[cl==2]<-c2
col[cl==3]<-c3
col[cl==4]<-c4

col1<-NULL
for (i in 1:length(a)){
    cola[i]<-col[a[i]]
    }
x
for (i in 1:length(b)){
    colb[i]<-col[b[i]]
    }
x
for (i in 1:length(c)){
    colc[i]<-col[c[i]]
    }
x
for (i in 1:length(d)){
    cold[i]<-col[d[i]]
    }
iniz<-rep(x0,4992)
fine<-rep(x1,4992)
```

```
segments(x0=iniz,x1=fine,y0=c(1:ngen),y1=c(1:ngen),col=c
(cola,colb,colc,cold))
}
```

A3.5.r8 `#Grafici sul raggruppamento dei geni in cluster (4 gruppi)`

```
clusterplot<-function (a,b,c,d,c1,c2,c3,c4,ngen)
{
x<-0
iniz<-rep(0,ngen)
fin<-rep(5,ngen)
plot(1,1,type="n",xlim=c(0,20),ylim=c(0,ngen),axes=F,xla
b="",ylab="")
axis(1,c(2.5,7.5,12.5,17.5),c("ALL-B","ALL-T","ALL-
Tr","AML"))

col1<-rep(c1,length(a))
col2<-rep(c2,length(b))
col3<-rep(c3,length(c))
col4<-rep(c4,length(d))
segments(x0=iniz,x1=fin,y0=c(1:ngen),y1=c(1:ngen),col=c(
col1,col2,col3,col4))
}
```

A3.5.r9 `#Grafici sul raggruppamento dei geni in cluster (3 gruppi)`

```
clusterpoint3<-function (a,b,c,cl,gruppo,c1,c2,c3,ngen)
{
if(gruppo==2){
  x0<-5
  x1<-10}
if(gruppo==3){
  x0<-10
  x1<-15}
if(gruppo==4){
  x0<-15
  x1<-20}
x<-0
iniz<-NULL
fine<-NULL
col<-NULL
cola<-NULL
colb<-NULL
colc<-NULL
col[cl==1]<-c1
```

A3.5.r10 `#Grafici sul raggruppamento dei geni in cluster (3 gruppi)`

```
clusterplot3<-function (a,b,c,c1,c2,c3,ngen)
{
```

```
x<-0
iniz<-rep(0,ngen)
fin<-rep(5,ngen)
plot(1,1,type="n",xlim=c(0,20),ylim=c(0,ngen),axes=F,xla
b="",ylab="")
axis(1,c(2.5,7.5,12.5,17.5),c("ALL-B","ALL-T","ALL-
Tr","AML"))

col1<-rep(c1,length(a))
col2<-rep(c2,length(b))
col3<-rep(c3,length(c))
segments(x0=iniz,x1=fin,y0=c(1:ngen),y1=c(1:ngen),col=c(
col1,col2,col3))

}
```

Altre *routines* utilizzate e già disponibili nelle varie librerie

Tipo di analisi	**Librerie**	**Funzioni**
Analisi *cluster*	`mva`	`hclust`
		`Dist`
Metodo delle *k*-medie	`mva`	`kmeans`
Metodo dei medoidi	*`cluster`*	`pam`

Capitolo 4
Analisi discriminante

4.1 Introduzione

Nel presente capitolo verrà affrontato il problema della costruzione di una regola discriminante per le quattro patologie. A tal fine verranno confrontate tre tecniche di analisi discriminante allo scopo di individuare quella più appropriata, di valutare l'efficacia dei tre metodi nella gestione un *dataset* avente un numero di variabili superiore al numero di osservazioni e di verificare se sia possibile selezionare un ridotto numero di variabili in grado fornire, mediante queste tecniche, una corretta classificazione delle unità.

In questa fase si è scelto di fare uso di variabili standardizzate e si sono fissate delle probabilità a priori per ciascun gruppo. È noto infatti che la leucemia linfoblastica acuta (ALL) si presenta nel 75% dei casi, mentre la forma mieloide acuta (AML) riguarda il 15% dei pazienti (il restante 10% è costituito da altre forme). Se si considera come evento totale l'appartenenza al gruppo ALL o AML, la probabilità a priori per le due classi è rispettivamente del 83,3% e 16,7%. All'interno del gruppo ALL si è poi equi ripartito questo 83% nelle tre sottoclassi (27.7% ciascuna).

4.2 Alberi di classificazione

Gli *alberi di classificazione* (si veda Breiman, L., Friedman, J., Olshen, R. and Stone, C. rif [53] oppure Hyunjoong Kim and Wei-Yin Loh rif [54]) sono una tecnica non parametrica, detta anche di *segmentazione gerarchica*, che

viene usata per prevedere la classe di appartenenza (variabile risposta) di un'unità statistica sulla base delle sue caratteristiche rappresentate delle variabili osservate (variabili esplicative). La regola discriminante è costituita da un sistema di condizioni rappresentato mediante una struttura ad albero in cui si distinguono nodi intermedi, nodi terminali, rami e radice. I nodi intermedi sono caratterizzati da una condizione riguardante una variabile esplicativa. I rami rappresentano il soddisfacimento o meno della condizione del nodo da cui discendono, i nodi terminali (o foglie) definiscono le classi associate alla regola rappresentata dal percorso che dalla radice porta ad essi. La flessibilità degli alberi di classificazione rende questo strumento uno dei più apprezzati nell'ambito dell'analisi di classificazione. Una caratteristica piuttosto interessante degli alberi di classificazione è quella di creare un ordinamento delle variabili secondo il loro potere discriminante: quanto più un nodo si trova vicino alla radice, tanto più la dicotomizzazione della variabile sulla quale si basa la condizione da verificare in quel nodo sarà rilevante ai fini della classificazione.

In prima battuta si sono quindi costruiti gli alberi di classificazione allo scopo di stilare una sorta di classifica per le variabili secondo il loro potere discriminante. Per problemi di carattere computazionale non è stato possibile utilizzare il *dataset* completo, si è pertanto fatto uso di quello ridotto.

La Figura 4.1 rappresenta l'albero di classificazione costruito sulla base di tutti ventidue i pazienti. Come si può notare, la regola fa uso di sole tre variabili su 218 che ha a disposizione, generando un tasso d'errore apparente nullo. I geni che utilizzati sono il numero 1180 (codice 2-013C04), 3156 (codice 2-033G12) e 9 (codice 2-001A09)

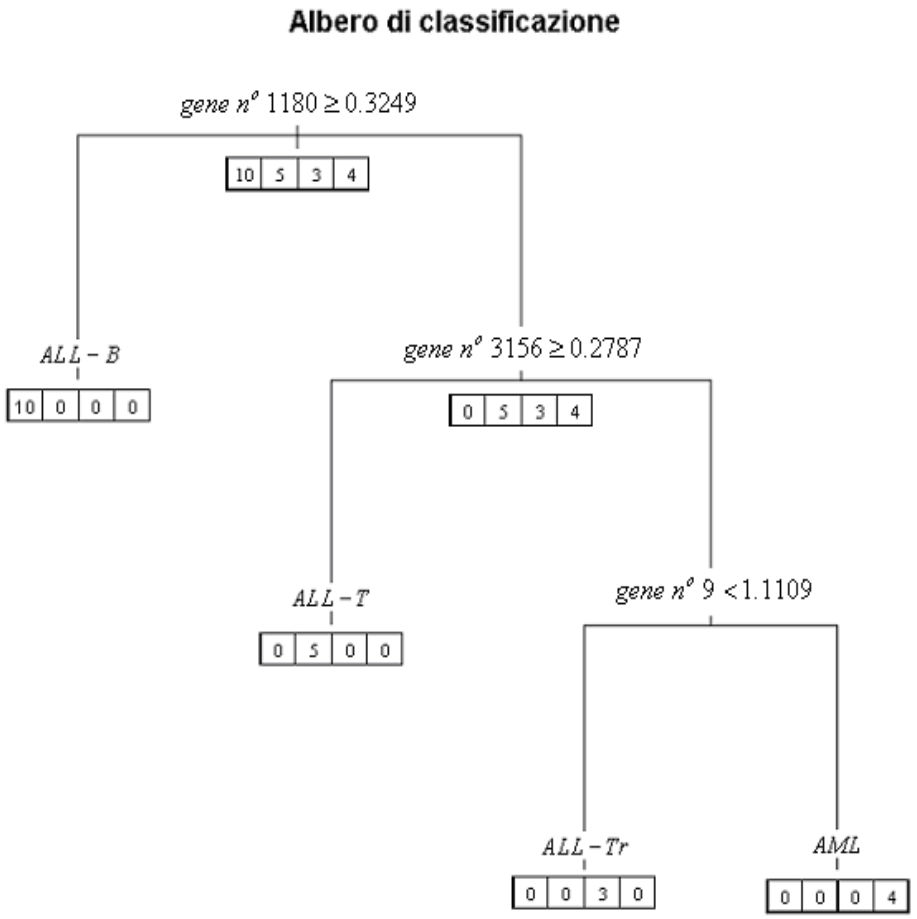

Figura 4.1 Albero di classificazione costruito con l'uso di tutte le variabili del dataset ridotto. In corrispondenza di ciascun nodo viene riepilogato, in una piccola tabella, il numero di unità presenti nel nodo distinte secondo la categoria cui appartengono. Ogni cella corrisponde ad una forma della patologia; nell'ordine si ha ALL-B, ALL-T, ALL-Tr, AML. I nodi intermedi e la radice riportano anche la condizione da verificare, mentre nei nodi terminali è indicata la classe da esso definita.

È noto però che il tasso d'errore apparente costituisce una sottostima del vero tasso di errore, in quanto la regola di discriminazione viene costruita proprio sulle unità da classificare; risulterà quindi ottima sull'insieme di dati analizzato ma in generale su campioni diversi non funziona così bene. Si è pertanto applicata una validazione incrociata di tipo *leave-one-out* che consiste

nell'eliminazione di un'unità per volta dal *dataset* e nella costruzione della regola sulle restanti *n-1* unità; successivamente l'osservazione eliminata viene classificata sulla base della regola di allocazione appena costruita. L'algoritmo viene iterato per *n* volte, in modo che ciascun soggetto venga classificato dalla regola costruita sulla base delle restanti *n-1* unità.

La Tabella 4.1 rappresenta la matrice di confusione ottenuta applicando la validazione di tipo *leave-one-out* sugli alberi di classificazione. In questo caso è molto evidente la sottostima del tasso d'errore apparente: la regola di allocazione infatti, non sembra funzionare così bene sulle nuove unità.

Alberi di classificazione		Classe prevista			
		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	8	1	1	0
	ALL-T	1	3	1	0
	ALL-Tr	1	1	0	1
	AML	2	0	0	2
Tasso d'errore:		0,409			

Tabella 4.1 Matrice di confusione relativa alla validazione incrociata di tipo *leave-one-out*

Come ricordato in precedenza, però, l'interesse nei confronti di questa tecnica, non è unicamente limitato alla creazione di una regola in grado di allocare le unità con un tasso d'errore più basso possibile, ma anche all'identificazione di variabili dotate di un buon potere discriminante. Nella Tabella 4.2 vengono quindi riportati i geni selezionati come *split* nei ventidue alberi di classificazione costruiti per la validazione incrociata di tipo *leave-one-out*, con il connesso codice identificativo ed il numero di volte in cui sono comparsi in questi ventidue alberi. Si noti che due di essi (il numero 9 e1180) sono stati usati per due volte nello stesso albero, a livelli diversi. Questa tabella tornerà utile nei prossimi paragrafi quando si parlerà di *analisi discriminante classica* e *logistica*.

Gene selezionato	Codice dientificativo	Frequenza
9	2-001A09	28
1180	2-013C04	17
134	2-002D02	15
21	2-001B09	5
898	2-010C10	1
981	2-011B09	1
848	2-009G08	1
3467	2-037A11	1
891	2-010C03	1

Tabella 4.2 Tabella riassuntiva dei geni selezionati come *split* nella validazione incrociata di tipo *leave-one-out*. Oltre al codice identificativo dei geni, si riportano le frequenze con le quali questi ultimi si sono presentati nei vari alberi di classificazione costruiti per la validazione incrociata.

In seconda battuta si è pensato di vagliare la capacità dagli alberi di classificazione nell'allocare le unità in sole due classi. Il passo successivo è stato dunque quello di definire quattro variabili dicotomiche $(y_{ALL\text{-}B}, y_{ALL\text{-}T}, y_{ALL\text{-}Tr}, y_{AML})$ che indicassero l'appartenenza o meno ad un certo gruppo: ad esempio per la patologia ALL-B si ha $y_{ALL\text{-}B}(x) = I_{ALL\text{-}B}(x)$, dove $I_{ALL-B}(x)$ è la funzione indicatrice che vale 1 se x appartiene al gruppo ALL-B e 0 altrimenti. Per ciascuna di queste variabili risposta è stato derivato il relativo albero di classificazione. Ora la regola fa uso di una sola variabile, ossia tutti gli alberi di classificazione sono dotati di un solo *split*. Le variabili selezionate in quest'unico *split* sono il gene numero 1180 (codice 2-013C04) per il gruppo ALL-B, il numero 487 (codice 2-015F09) per il gruppo ALL-T, il numero 3257 (codice 2-034H05) per il gruppo ALL-Tr ed il 1413 (codice 2-015F09) per il gruppo AML. In appendice si propongono i diagrammi di

dispersione di questi geni presi due a due nelle sei possibili combinazioni . Il tasso di errore apparente per ciascun albero è, anche questa volta, pari a zero.

Pure in questo caso è stata eseguita una validazione di tipo *leave-one-out*; di seguito vengono fornite le matrici di confusione ed i relativi tassi di errore per ciascuna patologia (Tabella 4.3). Si noti che il tasso di errore più elevato si ha nella patologia ALL-Tr, quella per cui si dispone di meno osservazioni; stupisce la buona precisione nella classe ALL-T che evidentemente risulta ben diversificata rispetto alle altre; piuttosto alto invece il tasso d'errore per la forma di leucemia ALL-B. Si noti infine che separare AML da "non AML" significa distinguere tra leucemia linfoide e mieloide. In Tabella 4.4 vengono forniti per ogni patologia, i geni selezionati per l'unico *split* evidenziato nei vari alberi costruiti per la validazione incrociata. Per ogni soggetto eliminato nella validazione, si sono evidenziati quattro geni in quanto si è costruito un albero per ciascuna delle quattro variabili $y_{ALL\text{-}B}, y_{ALL\text{-}T}, y_{ALL\text{-}Tr}, y_{AML}$; la parte a destra della tabella fornisce il codice identificativo di ciascun gene. La Tabella 4.5 illustra invece la frequenza con cui ogni gene è stato selezionato nella validazione incrociata di tipo *leave-one-out* anche questa tornerà utile nei prossimi paragrafi quando si parlerà di analisi discriminante classica e logistica.

Nel seguito ci si riferirà a quest'ultima analisi con il termine *alberi di classificazione per variabili risposta dicotomiche* per distinguerli da quella riferita ai primi alberi di classificazione che hanno fatto uso di variabile risposta politomica. Si noti che queste tecniche hanno selezionato alcuni geni in comune ossia i geni numero 1180, 9 e 898 e che i primi due compaiono in entrambi i casi con frequenza piuttosto alta.

Alberi di classificazione								
P: presente A: assente		Patologia						
		ALL-B		ALL-T		ALL-Tr		AML
		Classe prevista						
		P	A	P	A	P	A	P A
Classe reale	P	8	2	3	2	1	2	2 2
	A	3	9	0	17	6	13	1 17
Tasso d'errore		0,23		0,09		0,36		0,14

Tabella 4.3 Matrici di confusione relative agli alberi di classificazione. Per ciascuna patologia la variabile risposta è dicotomica ed individua la presenza o meno della patologia.

Patologia	Soggetto eliminato	Gene selezionato				Codice identificativo del gene			
		ALL-B	ALL-T	ALL-Tr	AML	ALL-B	ALL-T	ALL-Tr	AML
ALL-B	1	1180	487	3257	9	2-013C04	2-006A07	2-034H05	2-001A09
ALL-B	2	1180	487	3257	9	2-013C04	2-006A07	2-034H05	2-001A09
ALL-B	3	898	487	3257	9	2-010C10	2-006A07	2-034H05	2-001A09
ALL-B	4	1180	487	3257	9	2-013C04	2-006A07	2-034H05	2-001A09
ALL-B	5	1180	487	3257	9	2-013C04	2-006A07	2-034H05	2-001A09
ALL-B	6	1180	487	3257	9	2-013C04	2-006A07	2-034H05	2-001A09
ALL-B	7	1180	487	3257	9	2-013C04	2-006A07	2-034H05	2-001A09
ALL-B	8	1180	487	3257	9	2-013C04	2-006A07	2-034H05	2-001A09
ALL-B	9	1180	487	3257	9	2-013C04	2-006A07	2-034H05	2-001A09
ALL-B	10	1180	487	3257	9	2-013C04	2-006A07	2-034H05	2-001A09
ALL-T	11	1180	539	3257	9	2-013C04	2-006E11	2-034H05	2-001A09
ALL-T	12	1180	539	3257	9	2-013C04	2-006E11	2-034H05	2-001A09
ALL-T	13	1180	539	3257	9	2-013C04	2-006E11	2-034H05	2-001A09
ALL-T	14	1180	137	3257	9	2-013C04	2-002D05	2-034H05	2-001A09
ALL-T	15	1180	539	3257	9	2-013C04	2-006E11	2-034H05	2-001A09
ALL-Tr	16	1180	487	2317	9	2-013C04	2-006A07	2-025B01	2-001A09
ALL-Tr	17	1180	487	221	9	2-013C04	2-006A07	2-003C05	2-001A09
ALL-Tr	18	1180	487	103	9	2-013C04	2-006A07	2-002A07	2-001A09
AML	19	1180	487	2849	1324	2-013C04	2-006A07	2-030F05	2-014G04
AML	20	1180	487	1659	1413	2-013C04	2-006A07	2-018C03	2-015F09
AML	21	1180	487	1183	1413	2-013C04	2-006A07	2-013C07	2-015F09
AML	22	1180	487	487	1413	2-013C04	2-006A07	2-006A07	2-015F09

Tabella 4.4 Geni selezionati come unico *split* negli alberi di classificazione per variabili risposta dicotomiche. Per ogni soggetto eliminato nella validazione incrociata si sono costruiti quattro alberi, uno per ciascuna delle quattro variabili dicotomiche che individuano la patologia. La parte a destra della tabella fornisce il codice identificativo del gene.

Gene selezionato	Codice dientificativo	Frequenza
1180	2-013C04	18
487	2-006A07	18
9	2-001A09	18
3257	2-034H05	14
539	2-006E11	4
1413	2-015F09	3
981	2-011B09	1
898	2-010C10	1
891	2-010C03	1
848	2-009G08	1
137	2-002D05	1
3134	2-033F02	1
2317	2-025B01	1
221	2-003C05	1
103	2-002A07	1
2849	2-030F05	1
1695	2-018F03	1
1183	2-013C07	1
1324	2-014G04	1

Tabella 4.5 Frequenza con cui ogni gene viene selezionato come *split* nella validazione incrociata di tipo *leave-one-out*.

Per visualizzare il potere discriminante dei geni selezionati, si propongono le figure 4.2 e 4.3 che riportano le curve di Andrews per i soggetti. La prima utilizza i nove geni selezionati con la validazione incrociata di tipo *leave-one-out* sugli alberi di classificazione con variabile risposta politomica. La figura 4.3 utilizza i diciannove geni selezionati come *split* nella validazione incrociata sugli alberi di classificazione per variabili risposta dicotomiche. Si noti che in entrambi i casi, i gruppi che risultano meglio distinti sono l'ALL-B e l'ALL-T.

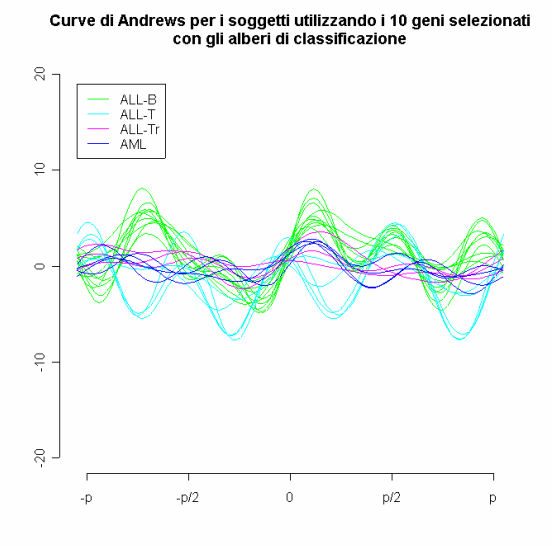

Figura 4.2 Curve di Andrews sui soggetti, costruite utilizzando i nove geni selezionati come *split* nella *cross-validation* di tipo *leave-one-out* sugli alberi di classificazione per variabile risposta politomica.

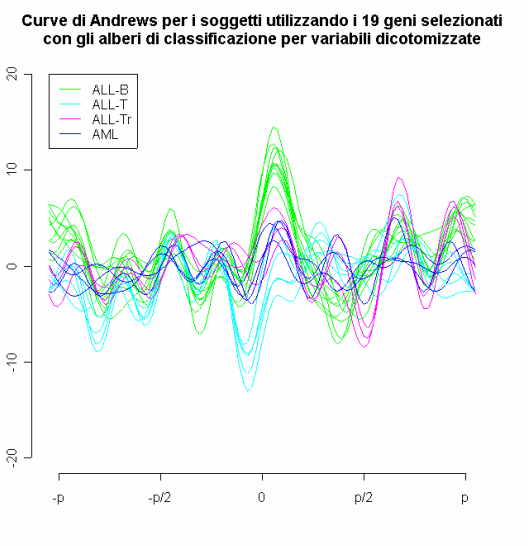

Figura 4.3 Curve di Andrews sui soggetti, costruite utilizzando i diciannove geni selezionati come *split* nella *cross-validation* di tipo *leave-one-out* sugli alberi di classificazione per variabili risposta dicotomiche.

Di seguito si propongono le curve di Andrews (Figura 4.5) ed i diagrammi a scatola (Figura 4.4) dei venticinque geni appena selezionati con la *cross-validation* di tipo *leave-one-out* sugli alberi di classificazione (nove con variabile politomica e diciannove per variabili dicotomiche, tre in comune) al fine di far risaltare eventuali differenze nel comportamento nella distribuzione delle espressioni da patologia a patologia.

Le curve di Andrews sono state tracciate per soli tre soggetti selezionati a caso da ciascun gruppo (tranne, ovviamente dal terzo che possiede solo tre unità) affinché le differenze, fossero imputabili solo a diversi livelli di espressione dei geni e non alle diverse numerosità dei gruppi. Si notano delle differenze marcate soprattutto tra il gruppo AML e tutti gli altri. Inoltre i profili del gruppo ALL-B stanno prevalentemente nella parte positiva dell'asse della

ordinate. Come fatto presente nel capitolo 3, questa è una caratteristica tipica dei geni molto espressi.

I diagrammi a scatola, distinti secondo la forma di leucemia, mettono in luce che alcuni geni presentano delle distribuzioni diverse a seconda dalla patologia dalla quale proviene la loro espressione.

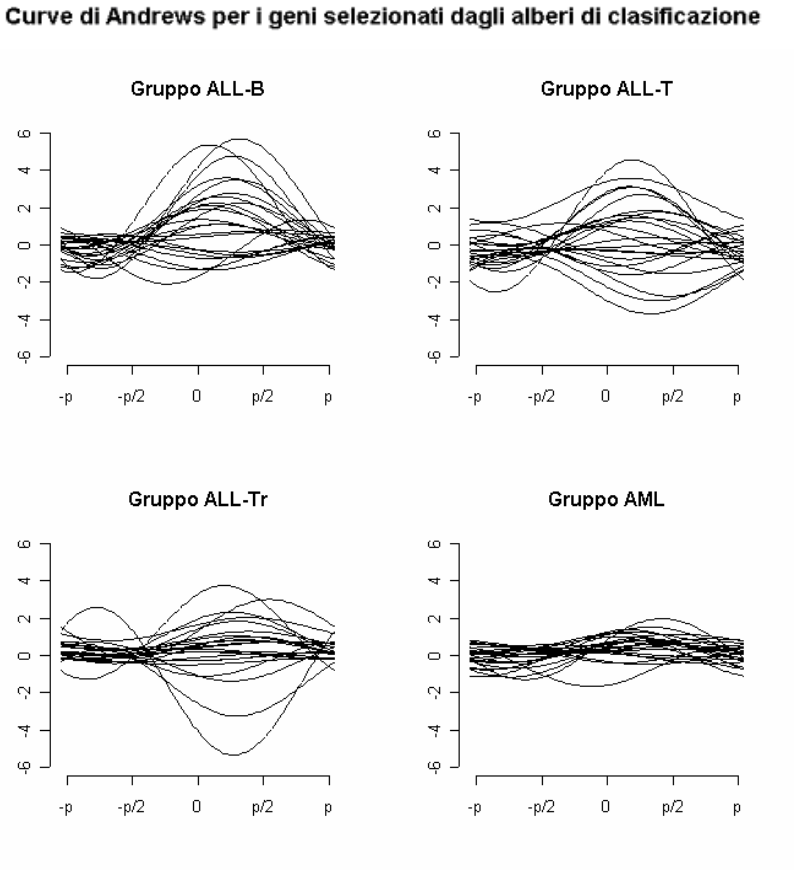

Figura 4.4 Curve di Andrews di tutti i geni selezionati dagli alberi di classificazione (sia quelli per variabili risposta dicotomiche che quelli su variabili politomiche).

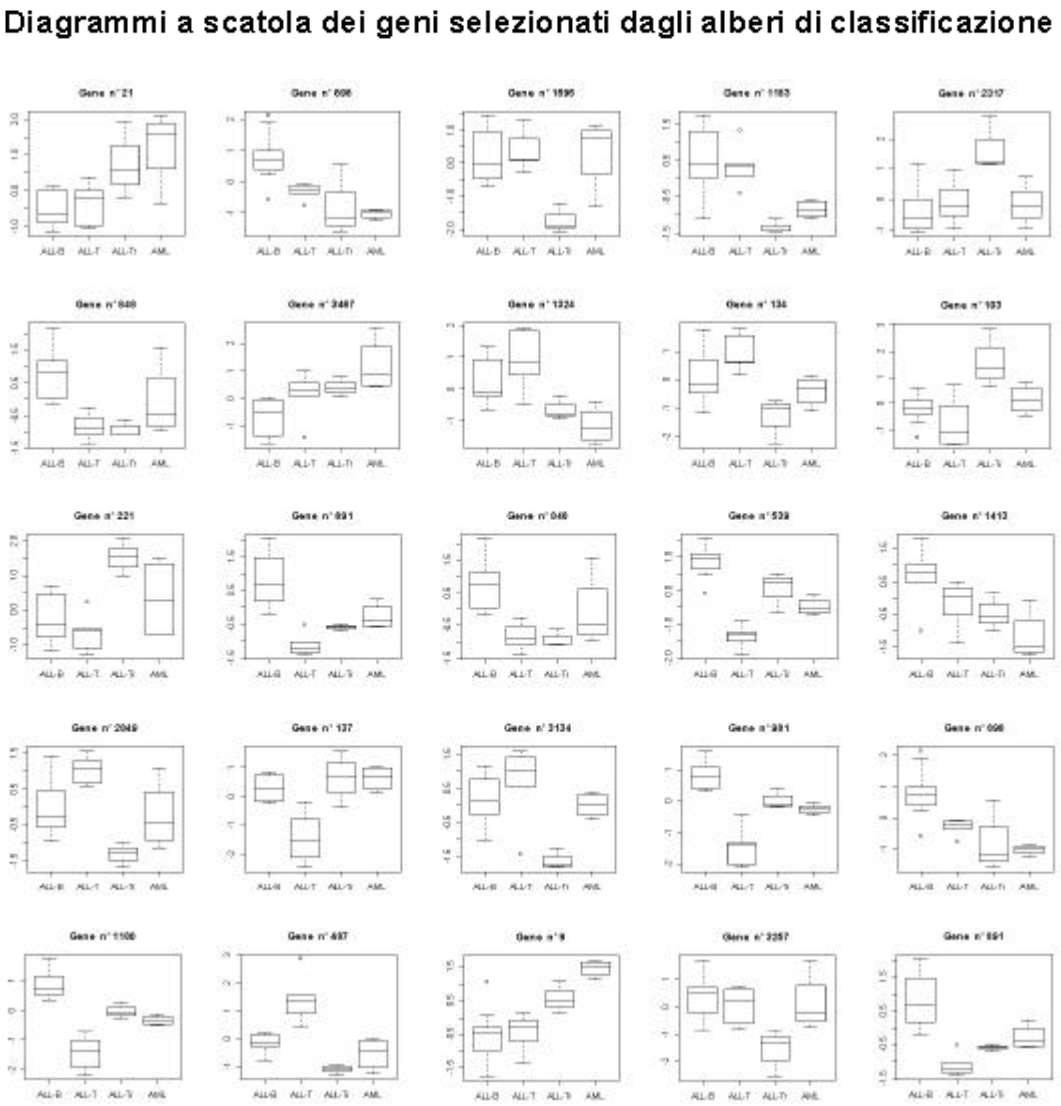

Figura 4.5 Box-plots dei 25 geni selezionati dagli alberi di classificazione con la validazione incrociata di tipo *leave-one-out*. Ciascun grafico riporta i box-plot delle espressioni di un gene distinti per le quattro forme di leucemia.

4.3 Analisi discriminante classica

Nel presente paragrafo si affronterà il problema di classificazione delle unità nei quattro gruppi di patologia mediante l'analisi discriminante classica (si veda Nagendra Kumar and Andreas G. Andreou rif [55]). L'utilizzo di tale tecnica sarà coadiuvato anche dalle informazioni provenienti dalle precedenti analisi sugli alberi di classificazione. Infatti, dal momento che l'utilizzo di un numero di variabili maggiore del numero di osservazioni è sconsigliato per questo tipo di analisi, i geni selezionati mediante gli alberi di classificazione saranno utilizzati come variabili riposta per l'analisi discriminante classica.

Introdotta da Fisher nel 1936, questa tecnica discriminante di tipo parametrico, si fonda sull'ipotesi di normalità della distribuzione dei caratteri nella popolazione. Si può distinguere tra *analisi discriminante lineare* e *analisi discriminante quadratica*; la prima si basa sull'assunto di omoschedasticità delle popolazioni, la seconda ammette invece che i gruppi abbiano una diversa matrice di varianze e covarianze (popolazioni eteroschedastiche). Avendo a disposizione un numero ridotto di osservazioni in ciascun gruppo, non è possibile, nel caso in esame, applicare l'analisi discriminante quadratica in quanto il numero di unità risulta insufficiente per la stima di quattro matrici di varianze e covarianze. D'ora in avanti ci si riferirà pertanto solo all'analisi discriminante lineare.

L'obiettivo dell'analisi discriminante è quello allocare l'unità *i* ad uno dei *G* gruppi in cui risulta suddivisa la popolazione sulla base del suo vettore *p*-dimensionale di osservazioni, x_i , in modo da commettere il minor numero possibile di errori di classificazione. Questo fine viene raggiunto attraverso una *regola di discriminazione* ossia una partizione di $\mathbb{R}^p$ in G regioni disgiunte: $\mathbb{R}_1,\ldots,\mathbb{R}_G$ taliche $\bigcup_{g=1}^{G}\mathbb{R}_g = \mathbb{R}^p$ mediante le quali si definisce una regola che alloca l'unità *i* alla popolazione *g* se $x_i \in \mathbb{R}_g$.

Con l'analisi discriminante lineare si cerca la combinazione lineare $z_j = \mathrm{X}\, a_j$ che separi al meglio i gruppi ossia che massimizzi la variabilità tra i gruppi rispetto a quella entro i gruppi. Ciò equivale a massimizzare il rapporto V_j tra forme quadratiche definito come segue

$$V_j = \frac{a_j' B a_j}{a_j' W a_j} \qquad j = 1 \ldots min(p, g-1)$$

dove B e W denotano le matrici di varianze e covarianze delle variabili rispettivamente tra ed entro i gruppi. I vettori a_i che massimizzano tale rapporto si ottengono dal calcolo degli autovettori dalla matrice $W^{-1}B$. Le combinazioni lineari $z_j = X\, a_j$ prendono il nome di *funzioni discriminanti* o *variabili canoniche*. Se si indica con $\bar{x}_g$ la media campionaria delle osservazioni nel gruppo *g*, la regola discriminante suggerisce di allocare l'unità *i* alla popolazione *g* se e solo se

$$\sum_{j=1}^{s} (a_j' x_i - a_j' \bar{x}_g)^2 = \min_k \sum_{j=1}^{s} (a_j' x_i - a_j' \bar{x}_k)^2 \qquad s = min\,(p,\, g\text{-}1) \qquad (1)$$

Dal punto di vista geometrico, ciò può essere interpretato in termini di distanza di Mahalanobis, definita come:

$$\left\{ (x - \mu_g)^{\mathrm{T}} \sum^{-1} (x - \mu_g) \right\}^{1/2},$$

con μ_g media del gruppo *g* e $g = 1 \ldots G$. Allocare un vettore di osservazioni *x* al gruppo *g* tale per cui vale (1), equivale ad allocare *x* alla popolazione g avente distanza di Mahalanibis stimata più piccola.

Si noti che, se come nel caso in esame, il numero di variabili è superiore al numero di osservazioni, la stima della matrice di varianze e covarienze risulta singolare e pertanto non invertibile. In tal caso in luogo di $\hat{\Sigma}^{-1}$ sarà utilizzata l'inversa generalizzata o pseudo-inversa di Moore –Penrose della matrice $\hat{\Sigma}$.

Per ovvi problemi di tipo computazionale, non è possibile applicare l'analisi discriminante lineare al *dataset* completo, anche in quest'occasione si farà pertanto uso di quello ridotto; la Tabella 4.6 fornisce i risultati della validazione incrociata di tipo *leave-one-out*.

Geni utilizzati		Classe prevista			
tutti 218		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	0	0
	ALL-T	0	5	0	0
	ALL-Tr	0	0	3	0
	AML	0	0	0	4
Tasso d'errore:		0			

Tabella 4.6 Matrice di confusione per l'analisi discriminante classica con l'uso di tutti i 218 geni del *dataset* ridotto.

La regola di allocazione classifica tutte le unità senza commettere errori, ma ovviamente risulta poco parsimoniosa in quanto fa uso di 218 variabili; inoltre la routine con la quale si è realizzata l'analisi avvisa l'utente dell'esistenza di collinearità tra le variabili. Come ricordato prima, infatti, il programma di elaborazione si trova a dover invertire una matrice singolare. Di *default* viene pertanto sostituita, al posto dell'inversa di questa matrice (che non sarebbe calcolabile), l'inversa generalizzata e viene proposto un messaggio che informa l'utente dell'avvenuta operazione.

Il primo tentativo di riduzione delle variabili esplicative è stato mosso utilizzando le informazioni fornite dall'analisi condotta nel precedente paragrafo con gli alberi di classificazione, sia quelli costruiti sulla variabile risposta politomica che quelli per variabili risposta dicotomiche. I risultati che si presenteranno nel seguito, si riferiscono ad analisi sulle quali è stata condotta una validazione incrociata di tipo *leave-one-out*.

4.3.1 Analisi basata sulle informazioni provenienti dagli alberi di classificazione con variabile risposta politomica.

Anzitutto, si sono usate come variabili esplicative i tre geni selezionati dal primo albero di classificazione, quello relativo a tutti ventidue soggetti. La Tabella 4.7 riporta la matrice di confusione ed il relativo tasso d'errore, pari a 0,227. L'albero di classificazione sembra non aver selezionato le variabili più significative dal punto di vista discriminante.

Successivamente, la stessa analisi è stata condotta su tutti i geni messi in evidenza dalla validazione incrociata di tipo *leave-one-out*, ossia quei nove (cfr Tabella 4.2) che hanno costituito gli *split* dei ventidue alberi di classificazione costruiti per la *cross-validation*. La matrice di confusione viene riportata nella Tabella 4.8

In seguito la regola di allocazione è stata costruita utilizzando solo i geni selezionati più d'una volta con gli alberi di classificazione, quindi geni numero 1180, 134, 9 e 21 (codici 2-013C04, 2-002D02, 2-001A09, 2-001B09 rispettivamente). Anche questa volta la regola sbaglia ad allocare due sole unità (Tabella 4.9)

Infine si sono usati solo i geni selezionati con maggior frequenza ossia i geni 1180 (codice2-013C04), 134 (codice 2-002D02) e 9 (codice 2-001A09). Tale regola ammette nuovamente un tasso di errore di 0,091 anche se cambiano le unità classificate erroneamente (Tabella 4.10).

Geni utilizzati		Classe prevista			
1180, 9, 3156		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	1	0
	ALL-T	0	3	2	0
	ALL-Tr	0	0	2	1
	AML	0	0	1	3
Tasso d'errore:		0,227			

Tabella 4.7

geni utilizati		Classe prevista			
9 geni selezionati dagli alberi di classificazione		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	10	0	0	0
	ALL-T	0	5	0	0
	ALL-Tr	0	0	2	1
	AML	0	0	1	3
Tasso d'errore:		0,091			

Tabella 4.8

Geni utilizzati		Classe prevista			
1180, 9, 134, 21		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	1	0
	ALL-T	0	5	0	0
	ALL-Tr	0	0	3	0
	AML	0	1	0	3
Tasso d'errore:		0,091			

Tabella 4.9

Geni utilizzati		Classe prevista			
1180, 134, 9		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	1	0
	ALL-T	0	4	0	1
	ALL-Tr	0	0	3	0
	AML	0	0	0	4
Tasso d'errore:		0,091			

Tabella 4.10

Tabella 4.7 Matrice di confusione per l'analisi discriminante classica con l'uso dei geni selezionati come *split* degli alberi di classificazione costruiti con l'uso di tutte le unità (senza la validazione incrociata).

Tabella 4.8 Matrice di confusione per l'analisi discriminante classica con l'uso dei 9 geni selezionati come *split* degli alberi di classificazione.

Tabella 4.9 Matrice di confusione per l'analisi discriminante classica con l'uso dei geni selezionati come *split* con frequenza maggiore di uno.

Tabella 4.10 Matrice di confusione per l'analisi discriminante classica con l'uso dei geni selezionati come *split* con maggior frequenza.

4.3.2 Analisi basata sulle informazioni provenienti dagli alberi di classificazione per variabile risposta dicotomica.

In questo caso, per prima cosa, si sono usate come variabili esplicative quelle selezionate dai primi quattro alberi di classificazione ossia quelli che si sono costruiti utilizzando tutti i soggetti: i geni numero 1180, 487, 3257 e 1413, (codici 2-013C04, 2-015F09, 2-034H05, 2-015F09, rispettivamente).

La matrice di confusione viene riportata in Tabella 4.11 il tasso d'errore è ora pari a 0,045: la regola alloca erroneamente una sola osservazione ma si noti che si stanno usando solo quattro variabili, non più 218.

In seguito l'analisi discriminante lineare è stata condotta utilizzando come variabili esplicative tutti i geni messi in evidenza dalla validazione incrociata sugli alberi di classificazione per variabili risposta dicotomiche costruiti nel paragrafo precedente, ossia quei geni, in tutto 19 (cfr Tabella 4.4 e Tabella 4.5), che sono stati selezionati come unico *split* negli alberi di classificazione costruiti eliminando di volta in volta un soggetto. I risultati della validazione *leave-one–out* sono riportati nella Tabella 4.12: la regola sbaglia cinque volte su ventidue (tasso d'errore 0,23). L'introduzione di nuove variabili non ha portato ad un miglioramento della capacità predittiva della regola ma sembra aver introdotto più rumore.

Successivamente, la regola di allocazione è stata costruita utilizzando solo i geni selezionati più d'una volta con gli alberi di classificazione, quindi geni numero 9, 487, 539, 1180, 1413, 3257 (codici 2-001A09, 2-006A04, 2-006E11, 2-013C04, 2-015F09, 2-034H05 rispettivamente). Tale regola ammette un tasso di errore di 0,091 (Tabella 4.13).

Infine si sono usati solo i geni selezionati con maggior frequenza ossia i geni 9, 487, 1180, 3257 (codici 2-001A09, 2-006A04, 2-013C04, 2-034H05 rispettivamente). Il tasso di errore rimane pari a 0,091 anche se cambiano le unità classificate erroneamente (Tabella 4.14).

Geni utilizzati		Classe prevista			
487, 1180, 1413, 3257		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	10	0	0	0
	ALL-T	0	5	0	0
	ALL-Tr	0	0	3	0
	AML	0	0	1	3
Tasso d'errore:		0,045			

Tabella 4.11

Geni utilizzati		Classe prevista			
19 selezionati dagli alberi di classificazione		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	8	0	1	1
	ALL-T	0	5	0	0
	ALL-Tr	0	0	1	2
	AML	0	0	1	3
Tasso d'errore:		0,23			

Tabella 4.12

Geni utilizzati		Classe prevista			
1180, 539, 487, 3257, 1413, 9		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	10	0	0	0
	ALL-T	0	5	0	0
	ALL-Tr	0	0	2	1
	AML	0	0	1	3
Tasso d'errore:		0,091			

Tabella 4.13

Geni utilizzati		Classe prevista			
1180, 539, 487, 1413		ALL-B	ALL-T	ALL_Tr	AML
Classe reale	ALL-B	9	0	1	0
	ALL-T	0	5	0	0
	ALL-Tr	0	0	3	0
	AML	0	0	1	3
Tasso d'errore:		0,091			

Tabella 4.14

Tabella 4.11 Matrice di confusione per l'analisi discriminante classica con l'uso dei geni selezionati come *split* degli alberi di classificazione per variabile risposta dicotomica costruiti con l'uso di tutte le unità (senza la validazione incrociata)

Tabella 4.12: Matrice di confusione per l'analisi discriminante classica con l'uso dei 19 geni selezionati come *split* degli alberi di classificazione per variabile risposta dicotomica.

Tabella 4.13 Matrice di confusione per l'analisi discriminante classica con l'uso dei geni selezionati come *split* con frequenza maggiore di uno negli alberi di classificazione per variabile risposta dicotomica.

Tabella 4.14 Matrice di confusione per l'analisi discriminante classica con l'uso dei geni selezionati come *split* con maggior frequenza negli alberi di classificazione per variabile risposta dicotomica.

Complessivamente si può affermare che gli alberi di classificazione costituiscono un buon strumento di selezione di variabili. Quelli per variabili risposta dicotomiche, in particolare, sono stati in grado di individuare quattro geni che classificano le unità con buona precisione (cfr Tabella 4.11). In

definitiva risulta evidente che, dal punto di vista strettamente statistico, sono sufficienti quattro geni (ossia i geni numero 1180, 487, 3257 e 1413) per caratterizzare il tipo di leucemia di cui è affetto un paziente, in quanto con questi quattro geni la regola sbaglia poco di più che utilizzando tutti i 218 geni della selezione. Una classificazione con un limitato tasso d'errore è possibile anche con l'utilizzo di poche variabili. La costruzione di una regola di allocazione semplice, che si serve di funzioni lineare delle variabili, sembra essere adeguata per allocare le unità con sufficiente precisione.

La figura che segue (Figura 4.6) propone le curve di Andrews per i quattro geni in grado di classificare le unità con un solo errore. Anche in questo caso, per evitare che le differenze fossero attribuibili alle diverse numerosità dei gruppi, i profili sono stati tracciati per soli tre soggetti selezionati a caso da ciascun gruppo.

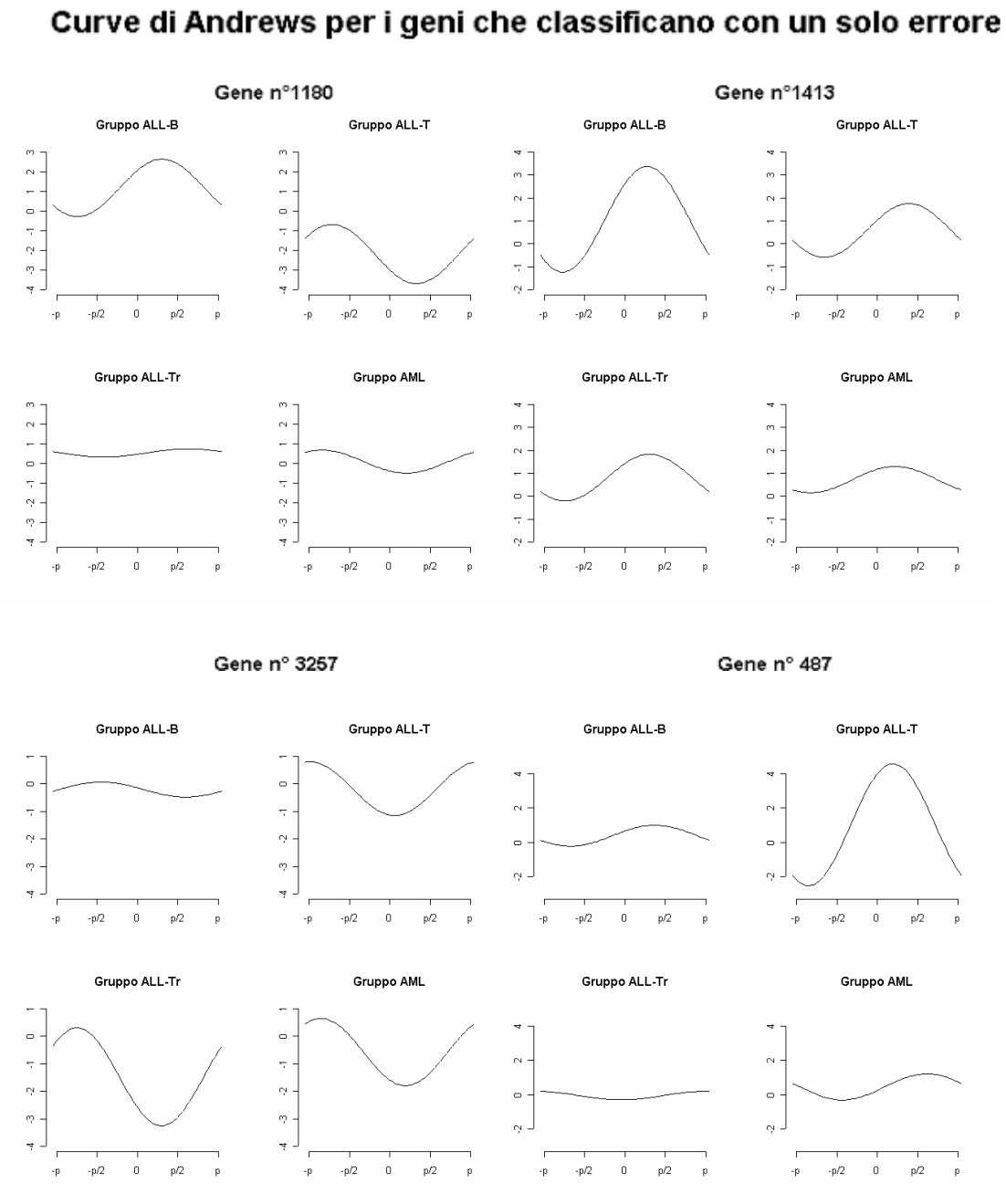

Figura 4.6 Curve di Andrews dei quattro geni che, attraverso l'analisi discriminante classica, allocano le osservazioni commettendo un solo errore di classificazione. Per ciascun

gene si sono utilizzate le osservazioni relative solo a tre soggetti in modo da rendere i grafici maggiormente confrontabili.

La Figura 4.7 riporta le curve di Andrews per i soggetti calcolate utilizzando i quattro geni che, con l'analisi discriminante lineare, classificano le unità con un solo errore. Si noti che, i gruppi che risultano meglio distinti sono l'ALL-B e l'ALL-T.

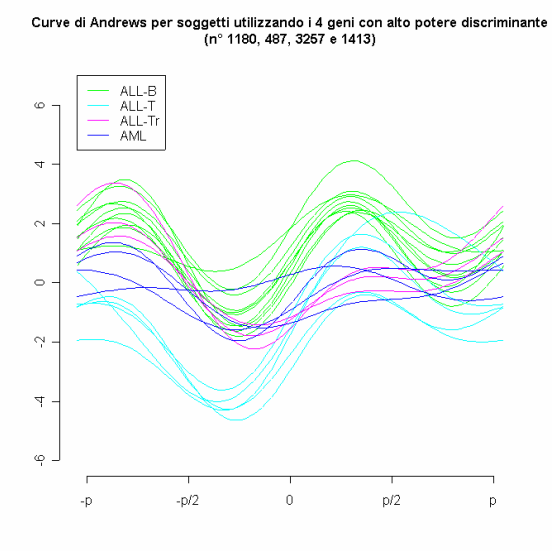

Figura 4.7 Curve di Andrews sui soggetti, costruite con l'utilizzo dei quattro che, con l'analisi discriminante lineare, classificano le unità con un solo errore.

4.4 Analisi discriminante logistica

Nel seguito verrà presentato l'utilizzo dell'analisi discriminante logistica finalizzato all'allocazione delle unità nelle quattro classi di patologia. Come nel precedente paragrafo, questa tecnica sarà integrata con l'utilizzo delle informazioni fornite dagli alberi di classificazione.

L'analisi *discriminante logistica* (Cox 1966) è un approccio di tipo semiparametrico, ossia non poggia su alcuna assunzione distributiva sui caratteri della popolazione. Tale approccio modella i logaritmi del rapporto tra le probabilità a posteriori di appartenenza alle popolazioni mediante espressioni del tipo:

$$log\left(\frac{P(G_i \mid x)}{P(G_j \mid x)}\right) = b_0 + b_1 f(x_1) + b_2 f(x_2) + ... + b_p f(x_p)$$

dove $x = (x_1, \ldots x_p)^T$, e $f_1, f_2 \ldots f_p$ sono trasformazioni delle variabili esplicative e $b_1, b_2 \ldots b_p$ sono parametri da stimare.

Oltre a non necessitare di assunti sulla distributivi, questa tecnica presenta il notevole vantaggio di poter essere applicata anche su variabili esplicative non continue. Inoltre, utilizzando la funzione di verosimiglianza per la stima dei coefficienti, è possibile disporre della loro distribuzione asintotica e quindi applicare le comuni tecniche inferenziali[3]. Ciò consente di applicare procedure iterative di selezione dei parametri.

Per condurre l'analisi logistica si è pertanto pensato di ripercorrere i passi effettuati per quella lineare. Inoltre, solo il caso in cui si sono utilizzate le variabili selezionate con gli alberi di classificazione per variabili risposta dicotomiche, si è applicata, ad ogni passo, la procedura iterativa di Akaike che consiste, partendo da un modello che fa uso di p le variabili, nell'eliminazione di quella associata al coefficiente $ß_i$ meno significativo e nella successiva ristima dei parametri, fino ad arrivare ad un modello che non può essere ulteriormente ridotto. Con i geni selezionati da questa procedura, si è riapplicata, ad ogni passo, la validazione incrociata di tipo *leave-one-out*.

La Tabella 4.15 riporta la matrice di confusione relativa all'applicazione dell'analisi discriminante logistica sul *dataset* ridotto. Con l'utilizzo dei 218 geni della selezione, la regola sbaglia ad allocare un soggetto della classe ALL-

[3] Nel caso in esame, l'utilizzo di tecniche inferenziali, e di riferimenti a risultati asintotici, sono in realtà un po'azzardati, in quanto la numerosità di cui si dispone è ben lungi dal rispecchiare le condizioni cui si fa riferimento.

B e uno della classe AML; si noti però che il soggetto della classe ALL-B è lo stesso che veniva classificato erroneamente in diverse occasioni quando è stato condotto il confronto tra i metodi di imputazione (cfr paragrafo 2.1.2)

In questo caso, inoltre, la procedura di Akaike suggerirebbe di lasciare inalterato il modello e di non eliminare nemmeno un variabile. Ciò è dovuto all'elevata dimensionalità del *dataset* che rende poco efficace questa tecnica poco efficace.

Geni utilizzati		Classe prevista			
tutti 218		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	0	0
	ALL-T	0	5	0	0
	ALL-Tr	1	0	3	1
	AML	0	0	0	3
Tasso d'errore:		0,091			

Tabella 4.15 Matrice di confusione per l'analisi discriminante logistica con l'uso di tutti i 218 geni del *dataset* ridotto.

Gli altri risultati vengono presentati nelle tabelle 4.15-4.22 quelle contrassegnate con la lettera (b) si riferiscono all'analisi condotta sui i geni selezionati mediante la procedura di Akaike dalla tabella (a) corrispondente (come detto poc'anzi la procedura di Akaike è stata applicata solo per i modelli costruiti sui geni selezionati con gli alberi di classificazione per variabili risposta dicotomiche).Il tasso di errore più basso che si riesce ad ottenere è dello 0,091.

❖ Geni selezionati dagli gli alberi di classificazione per variabile risposta politomica

Geni utilizzati		Classe prevista			
1180, 9, 3156		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	8	1	0	1
	ALL-T	0	5	0	0
	ALL-Tr	0	0	2	1
	AML	0	0	2	2
Tasso d'errore:		0,227			

Tabella 4.16

geni utilizzati		Classe prevista			
10 selezionati dagli alberi di classificazione		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	0	1
	ALL-T	0	5	0	0
	ALL-Tr	0	0	3	0
	AML	0	0	1	3
Tasso d'errore:		0,091			

Tabella 4.17

Geni utilizzati		Classe prevista			
1180, 9, 134, 21		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	1	0
	ALL-T	0	4	1	0
	ALL-Tr	1	0	2	0
	AML	0	0	1	3
Tasso d'errore:		0,182			

Tabella 4.18

Geni utilizzati		Classe prevista			
1180, 134, 9		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	8	0	1	1
	ALL-T	0	5	0	0
	ALL-Tr	1	0	2	0
	AML	0	0	1	3
Tasso d'errore:		0,182			

Tabella 4.19

Tabella 4.16 Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni selezionati per gli *split* degli alberi di classificazione per variabile risposta politomica costruiti con l'uso di tutte le unità (senza la validazione incrociata)

Tabella 4.17 Matrice di confusione per l'analisi discriminante logistica con l'uso dei 9 geni selezionati per gli *split* degli alberi di classificazione per variabile risposta politomica.

Tabella 4.18 Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni selezionati come *split* con frequenza maggiore di uno (per gli alberi di classificazione per variabile risposta politomica).

Tabella 4.19 Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni selezionati come *split* con maggior frequenza (per gli alberi di classificazione per variabile risposta politomica).

❖ Geni selezionati dagli gli alberi di classificazione per variabile risposta dicotomizzata

Geni utilizzati		Classe prevista			
487, 1180, 1413, 3257		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	1	0
	ALL-T	0	5	0	0
	ALL-Tr	0	0	3	0
	AML	0	0	1	3
Tasso d'errore:		0,091			

Tabella 4.20 (a)

Geni utilizzati		Classe prevista			
19 selezionati dagli alberi di classificazione		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	7	0	2	1
	ALL-T	0	5	0	0
	ALL-Tr	0	0	2	1
	AML	0	0	0	3
Tasso d'errore:		0,182			

Tabella 4.21 (a).

Geni utilizzati		Classe prevista			
1180, 539, 487, 3257, 1413, 9		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	0	0
	ALL-T	0	5	0	0
	ALL-Tr	1	0	3	1
	AML	0	0	0	3
Tasso d'errore:		0,091			

Tabella 4.22 (a)

Geni utilizzati		Classe prevista			
1180, 539, 487, 1413		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	0	0
	ALL-T	0	5	0	0
	ALL-Tr	1	0	3	1
	AML	0	0	0	3
Tasso d'errore:		0,091			

Tabella 4.23 (a)

Tabella 4.20 (a) Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni selezionati per gli *split* degli alberi di classificazione costruiti con l'uso di tutte le unità (senza la validazione incrociata)

Tabella 4.21 (a) Matrice di confusione per l'analisi discriminante logistica con l'uso dei 19 geni selezionati per gli *split* degli alberi di classificazione.

Tabella 4.22 (a) Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni selezionati come *split* con frequenza maggiore di uno.

Tabella 4.23 (a) Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni selezionati come *split* con maggior frequenza.

❖ Applicazione del metodo di Akaike

Geni utilizzati		Classe prevista			
1180, 1413		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	0	1
	ALL-T	0	4	0	1
	ALL-Tr	1	0	2	0
	AML	0	1	0	3
Tasso d'errore:		0,182			

Tabella 4.20 (b)

Geni utilizzati		Classe prevista			
487, 1413		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	0	1
	ALL-T	1	4	0	0
	ALL-Tr	0	0	2	1
	AML	0	0	1	3
Tasso d'errore:		0,182			

Tabella 4.21 (b)

Geni utilizzati		Classe prevista			
487, 1413		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	0	1
	ALL-T	1	4	0	0
	ALL-Tr	0	0	2	1
	AML	0	0	1	3
Tasso d'errore:		0,182			

Tabella 4.22 (b)

Geni utilizzati		Classe prevista			
487, 1413		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	9	0	0	1
	ALL-T	1	4	0	0
	ALL-Tr	0	0	2	1
	AML	0	0	1	3
Tasso d'errore:		0,182			

Tabella 4.23 (b)

Tabella 4.20 (b) Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni evidenziati con il criterio di Akaike dalle variabili selezionate per gli *split* degli alberi di classificazione costruiti con l'uso do tutte le unità (senza la validazione incrociata)

Tabella 4.21 (b) Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni evidenziati con il criterio di Akaike dalle 19 variabili selezionate per gli *split* degli alberi di classificazione

Tabella 4.22 (b) Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni selezionati dal criterio di Akaike delle variabili evidenziate come *split* con frequenza maggiore di uno

Tabella 4.23 (b) Matrice di confusione per l'analisi discriminante logistica con l'uso dei geni selezionati dal criterio di Akaike delle variabili evidenziate come *split* con frequenza maggiore

Se confrontati con il precedente paragrafo, i risultati mettono in luce che l'analisi discriminante lineare sembra più adeguata per trattare il dataset analizzato. Essa è infatti più parsimoniosa essendo in grado di produrre gli stessi tassi d'errore ma con l'utilizzo di un numero minore di variabili.

Le analisi discriminante classica e logistica, sono state condotte anche su alcuni dei geni che erano stati messi in evidenza dalle curve di Andrews uscenti per una sola patologia dalle bande fissate a diversi livelli (con il criterio percentuale oppure partendo da ±4 con passo di due cfr appendice tabelle in appendice al capitolo 3). I risultati hanno però fornito delle matrici di confusione con un tasso di errata classificazione piuttosto elevato. Probabilmente all'interno di questa selezione persiste ancora molto rumore.

Infine si sono usate come variabili esplicative i medoidi, sia quelli selezionati sia dal *dataset* completo che quelli selezionati dal *dataset* ridotto. Sebbene dai risultati (riportati in appendice nelle Tabella A.4.1 e Tabella A.4.2) si intuisce che i medoidi non siano particolarmente adatti per questo tipo di analisi, il tentativo compiuto è utile per mettere nuovamente in luce che le analisi forniscono risultati migliori se applicate al *dataset* ridotto. Trova quindi conferma l'ipotesi che nel *dataset* completo ci sia esubero si informazione che occulta i dati utili a fini discriminanti. C'è da chiedersi se anche nel *dataset* ridotto ci siano variabili che creano solo rumore senza contribuire in modo determinante nella classificazione. Questa è una domanda alla quale ci si propone di rispondere nel capitolo che segue.

Appendice

A4.1 Selezione dei geni nell'analisi discriminante

Diagrammi di dispersione relativi ai quattro geni selezionati come *split* dagli alberi di classificazione su variabili risposta dicotomiche costruiti utilizzando tutti 22 i soggetti ed in grado di classificare le unità con un solo errore mediante l'analisi discriminante lineare (paragrafo 4.3). I geni sono stati abbinati due a due nelle sei possibili combinazioni.

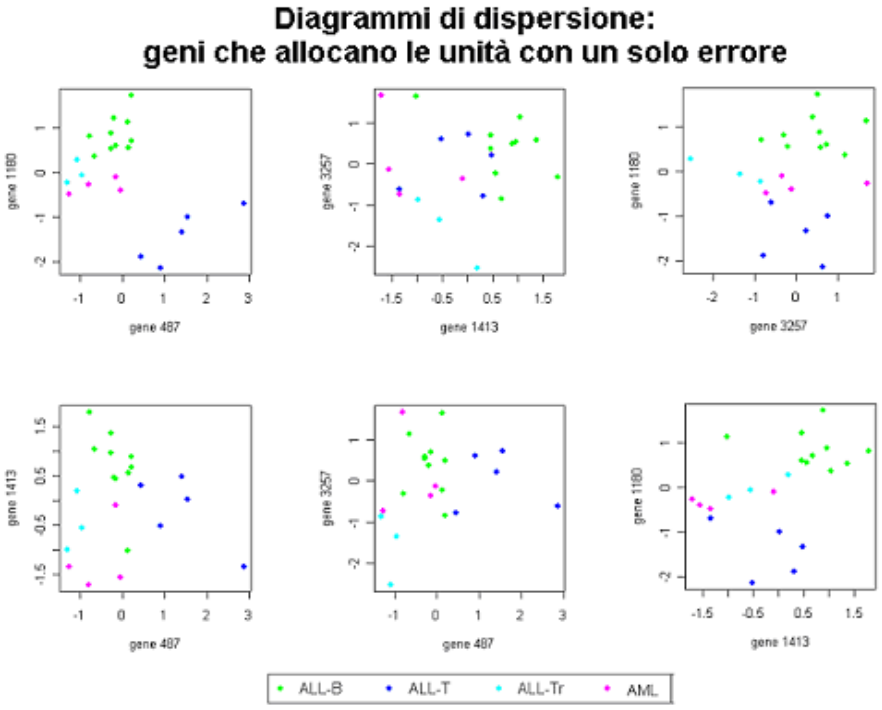

Figura A4.1 Diagrammi di dispersione dei quattro geni selezionati dagli alberi di classificazione senza la validazione incrociata. Ogni grafico raccoglie le espressioni delle sei possibili coppie di geni, i diversi colori distinguono i quattro tipi di leucemia

A4.2 Analisi discriminante classica e logistica sui medoidi

Risultati relativi all'applicazione dell'analisi discriminante classica e logistica utilizzando come variabili esplicative i geni selezionati come medoidi durante la fase di analisi esplorativa. Si sono utilizzati prima i medoidi della suddivisione in quatto gruppi e poi quelli delle suddivisione in tre ed infine quelli comuni alle due suddivisioni. Si noti che con i medoidi del il *dataset* ridotto si ottengono risultati migliori.

Analisi discriminante classica con i medoidi

(medoidi del dataset completo)

Geni utilizzati		Classe prevista			
Medoidi della suddivisione in tre gruppi		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	4	2	1	3
	ALL-T	3	0	1	1
	ALL-Tr	1	1	1	0
	AML	0	2	1	1
Tasso d'errore:		0,727			

Geni utilizzati		Classe prevista			
Medoidi della suddivisione in quattro gruppi		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	0	3	4	3
	ALL-T	2	0	2	1
	ALL-Tr	0	2	0	1
	AML	2	1	0	1
Tasso d'errore:		0,954			

Geni utilizzati		Classe prevista			
Medoidi comuni alle suddivisioni		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	6	2	2	0
	ALL-T	1	1	2	1
	ALL-Tr	1	2	0	0
	AML	1	2	1	0
Tasso d'errore:		0,682			

(medoidi del dataset ridotto)

Geni utilizzati		Classe prevista			
Medoidi della suddivisione in tre gruppi		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	8	0	0	2
	ALL-T	0	5	0	0
	ALL-Tr	0	0	2	1
	AML	0	0	1	3
Tasso d'errore:		0,182			

Geni utilizzati		Classe prevista			
Medoidi della suddivisione in quattro gruppi		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	8	1	0	1
	ALL-T	0	5	0	0
	ALL-Tr	0	0	3	0
	AML	0	0	0	4
Tasso d'errore:		0,091			

Geni utilizzati		Classe prevista			
Medoidi comuni alle suddivisioni		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	6	2	1	1
	ALL-T	0	3	1	1
	ALL-Tr	0	0	3	0
	AML	0	0	1	3
Tasso d'errore:		0,318			

Tabella A.4.1 Analisi discriminante classica sui medoidi

Analisi discriminante logistica con i medoidi

(medoidi del dataset completo)

Geni utilizzati		Classe prevista			
Medoidi della suddivisione in tre gruppi		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	5	2	1	2
	ALL-T	2	1	0	2
	ALL-Tr	0	1	0	2
	AML	2	1	1	0
Tasso d'errore:		0,727			

Geni utilizzati		Classe prevista			
Medoidi della suddivisione in quattro gruppi		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	1	4	1	4
	ALL-T	2	2	1	0
	ALL-Tr	1	2	0	0
	AML	3	0	1	0
Tasso d'errore:		0,864			

Geni utilizzati		Classe prevista			
Medoidi comuni alle suddivisioni		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	6	2	1	1
	ALL-T	1	0	2	2
	ALL-Tr	1	2	0	0
	AML	2	0	2	0
Tasso d'errore:		0,727			

(medoidi del dataset ridotto)

Geni utilizzati		Classe prevista			
Medoidi della suddivisione in tre gruppi		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	8	0	1	1
	ALL-T	0	4	0	1
	ALL-Tr	1	0	1	1
	AML	0	0	1	3
Tasso d'errore:		0,273			

Geni utilizzati		Classe prevista			
Medoidi della suddivisione in quattro gruppi		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	8	0	1	1
	ALL-T	0	5	0	0
	ALL-Tr	0	0	3	0
	AML	1	0	0	3
Tasso d'errore:		0,136			

Geni utilizzati		Classe prevista			
Medoidi comuni alle suddivisioni		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	7	1	1	1
	ALL-T	0	3	0	2
	ALL-Tr	1	1	1	0
	AML	0	1	1	2
Tasso d'errore:		0,409			

Tabella A.4.2 Analisi discriminante logistica sui medoidi

R*outines* utilizzate e già disponibili nelle varie librerie

Tipo di analisi	**`Libreria`**	**`Funzione`**
Alberi di classificazione	`rpart`	`rpart`
Analisi discriminante lineare	`MASS`	`lda`
Analisi discriminante quadratica	`MASS`	`qda`
Analisi discriminante logistica	nnet	`multinom`

Capitolo 5

Metodi per il controllo della dimensionalità

5.1 Introduzione

Nel corso del precedente capitolo si sono adottati i metodi convenzionali per l'analisi discriminante, affrontando il problema dell'elevata dimensionalità prevalentemente mediante la selezione di variabili dotate di buon potere discriminante. Tale selezione è stata effettuata manualmente con gli alberi di classificazione oppure con il metodo di Akaike ed ha fornito dei buoni risultati riuscendo ad isolare pochi geni con elevato potere discriminante. Si è avuto modo però di sottolineare che il principio della parsimonia non è altrettanto gradito in biologia, in quanto è vivo interesse dei ricercatori individuare gruppi, anche piuttosto numerosi, di geni responsabili della diffusione di una patologia o del suo differenziarsi in varie forme.

Nel presente capitolo si sperimenteranno metodi di selezione automatica e metodi di controllo della dimensionalità attraverso l'utilizzo di parametri di penalizzazione. Questo consentirà non solo di vagliare altre tecniche di classificazione in grado di gestire un numero di variabili superiore al numero di unità statistiche, ma anche la selezione di geni significativi dal punto di vista biologico.

5.2 Metodo " nearest shrunken centroid "

Recentemente proposto da Tibshirani *et al* rif [52], questo metodo consente non solo di costruire una regola di allocazione, ma anche di mettere in evidenza il gruppo di geni che dà il contributo maggiore per una corretta classificazione. Inoltre, a differenza dei precedenti metodi, non presenta particolari problemi computazionali e può pertanto essere applicato sul *dataset* completo, consentendo così un utile confronto tra i geni selezionati da questo algoritmo e quelli selezionati dai ricercatori (*dataset* ridotto).

Il metodo si basa sul criterio del centroide più vicino (*nearest centroid*) ma presenta la sostanziale differenza di utilizzare solo i centroidi di gruppo che più si discostano dalla media generale, ovvero quelli per cui la differenza dal centroide generale supera una certa soglia regolata da un parametro (Δ) fissato con la validazione incrociata.

Sia allora x_{ij} l'espressione del gene, $i = 1\ldots p$, nell'individuo, $j = 1\ldots n$, e si indichi con C_k la classe k composta da n_k elementi, $k = 1\ldots K$. L'i-esimo centroide della classe k sarà pertanto dato da $\bar{x}_{ik} = \sum_{j \in C_k} \frac{x_{ij}}{n_k}$ (media di gruppo) mentre il centroide generale sarà $\bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n}$ (media generale).

Si definisce

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k (s_i + s_0)} ,$$

dove s_i è la radice quadrata della somma delle varianze di classe per il gene i :

$$s_i^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 ,$$

e $m_k = \sqrt{\frac{1}{n_k} + \frac{1}{n}}$. Tale quantità è definita in modo che $m_k \cdot s_i$ sia pari alla stima dello standard error del numeratore di d_{ik} . Nel denominatore, s_0 è una costante (uguale per tutti i geni) pari alla mediana di s_i che viene introdotta per evitare

che geni con un basso livello di espressione generino valori elevati di d_{ik}. Si noti che d_{ik} è il valore della statistica *t* per il seguente sistema d'ipotesi:

$$\begin{cases} H_0: & m_{ik} = m_i \\ H_1: & m_{ik} \neq m_i \end{cases},$$

con m_{ik} media del gene *i* rispetto al il gruppo *k* e m_i media del gene *i* su tutti i gruppi.

Dall'espressione che definisce d_{ik} si ricava che

$$\overline{x}_{ik} = \overline{x} + m_k (s_i + s_0) \cdot d_{ik} \ .$$

Il metodo proposto da Tibshirani consiste nel confrontare d_{ik}, i= 1,…,*p e* k=1,…,*K*, con la soglia Δ producendo i coefficienti d'_{ik} definiti come segue

$$d'_{ik} = sign(d_{ik})\left(|\, d_{ik}\,| - \Delta\right)_+$$

dove $(\bullet)_+$ è la funzione *parte positiva* ($x_+ = x$ se $x > 0$ e 0 altrimenti).
Questi coefficienti servono poi per il calcolo degli "*shrunken centroids*" definiti come:

$$\overline{x}'_{ik} = \overline{x} + m_k (s_i + s_0) \cdot d'_{ik} \ .$$

È chiaro che, incrementando il parametro Δ, diminuisce il numero di geni coinvolti nella regola di allocazione. In particolare, se per il gene *i*, d'_{ik} viene fissato a zero per tutte le *K* classi, allora il centroide per il gene *i* è $\overline{x}_i$ per tutte le classi. Ciò significa che il gene *i* non contribuisce al calcolo della regola discriminante.

La regola consiste nell'allocare l'unità x^* alla classe *k* tale per cui la seguente quantità, denominata *punteggio discriminante*, è minima

$$d_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \overline{x}'_{ik})^2}{(s_i + s_0)^2} - 2\log p_k \ .$$

Si noti che il punteggio discriminante consente anche di tener conto delle probabilità a priori, che sono state fissate come nelle analisi proposte in precedenza.

Questa tecnica è stata applicata al *dataset* completo ed a quello ridotto, su variabili standardizzate e non. Di seguito vengono riportati solamente i risultati relativi all'applicazione dell'algoritmo su variabili standardizzate; per gli altri risultati si rimanda all'appendice (tabelle A5.1 e A5.2, figure A5.1 e A5.2).
La Figura 5.1 mostra le prestazioni dell'algoritmo al variare del coefficiente Δ: il tasso d'errore minimo (0,091) si ottiene per valori di Δ compresi tra 1,3 e 1,4.

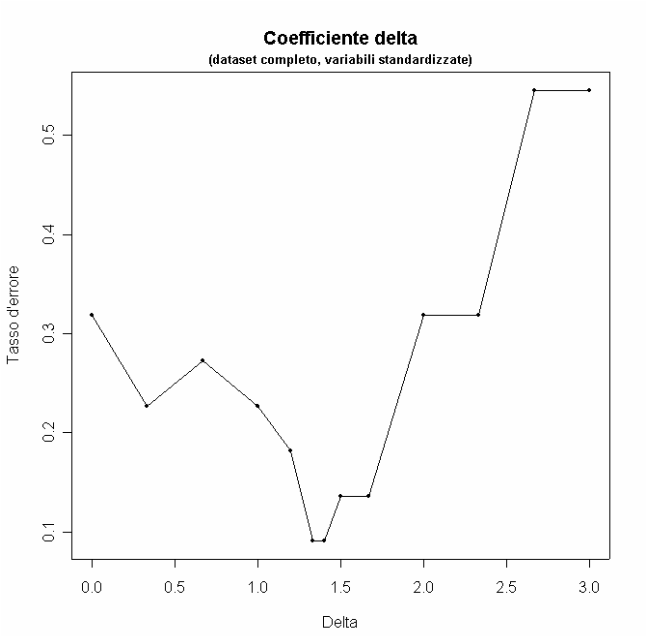

Figura 5.1 Efficacia dell'algoritmo applicato sul *dataset* ridotto con varabili standardizzate: tasso d'errore al variare del coefficiente Δ.

Alla luce di ciò si è deciso di fissare $\Delta = 1,3$. La matrice di confusione viene riportata in Tabella 5.1; si noti che le due osservazioni classificate erroneamente appartengono al gruppo in cui si dispone di sole tre osservazioni; ma la cosa più interessante è senza dubbio il controllo dei geni selezionati. La Tabella 5.2 ne fornisce una panoramica evidenziando quelli comuni al *dataset* ridotto e di quelli diversi da quest'ultimo. Si noti che generalmente in ogni patologia vengono selezionati dei geni diversi rispetto a quelli individuati nelle altre (ossia sono pochi i geni selezionati in più d'una patologia).

Metodo shrunken centroids		Classe prevista			
		ALL-B	ALL-T	ALL-Tr	AML
Classe reale	ALL-B	10	0	0	0
	ALL-T	0	5	0	0
	ALL-Tr	2	0	1	0
	AML	0	0	0	4
Tasso d'errore:		0,091			

Tabella 5.1 Matrice di confusione relativa all'applicazione del metodo *shrunken centroid* con la validazione incrociata di tipo *leave-one-out*

Variabili standardizzate, dataset completo		
Geni selezionati dalla procedura shrunken centroids	ALL-B	539 891 981 1180 2481 3209 3810 4002 4060 4195 4241
	ALL-T	137 283 400 487 535 539 570 611 679 729 853 880 884 891 895 952 981 1009 1180 1364 1409 1455 1492 1549 1650 1711 1712 1714 1824 2034 2051 2066 2068 2134 2218 2254 2272 2433 2441 2458 2506 2570 2673 2697 2710 2732 2748 2868 3110 3124 3156 3253 3418 3601 3787 3810 3828 4002 4021 4060 4103 4142 4178 4195 4241 4275 4420 4480 4489 4697
	ALLTr	57 103 221 404 601 939 1064 1378 1553 1673 1677 1695 1710 1726 1730 2277 2317 2743 3134 3257 3483 3683 3841 3846 3862 3869 3882 3987 4089 4112 4174
	AML	9 21 47 191 295 348 363 421 452 474 558 577 621 630 664 674 731 922 948 1025 1113 1282 1471 1495 1530 1596 1604 1643 1711 1781 1831 1877 1885 2016 2019 2044 2046 2052 2103 2119 2121 2164 2165 2200 2461 2466 2486 2818 2848 2964 3105 3216 3246 3312 3441 3576 3659 3685 3691 3717 3837 3864 3911 3967 4107 4110 4122 4179 4236 4237 4281 4312 4445 4473 4508 4525 4664 4681 4697 4762 4988
Geni selezionati e comuni al dataset ridotto	ALL-B	539 891 981 1180 2481 3209 3810 4002 4060 4195 4241
	ALL-T	137 400 487 535 539 570 611 729 853 880 884 891 895 981 1180 1364 1409 1549 1650 1711 1712 1714 2034 2051 2218 2254 2433 2441 2458 2673 2697 2710 2732 2748 2868 3110 3124 3156 3253 3418 3601 3810 3828 4002 4021 4060 4103 4142 4178 4195 4241 4275 4420 4480 4489 4697
	ALL-Tr	103 221 1378 1553 1677 1695 1710 1730 2317 2743 3134 3257 3483 3683 3841 3862 3869 3987 4174
	AML	9 21 452 474 558 621 664 674 922 1025 1113 1282 1471 1530 1711 1831 2016 2019 2052 2103 2200 2461 2486 2848 2964 3246 3864 3967 4110 4236 4445 4473 4508 4525 4681 4697 4762
geni selezionati e non comuni al dataset ridotto	ALL-B	
	ALL-T	283 679 952 1009 1455 1492 1824 2066 2068 2134 2272 2506 2570 3787
	ALL-Tr	57 404 601 939 1064 1673 1726 2277 3846 3882 4089 4112
	AML	47 191 295 348 363 421 577 630 731 948 1495 1596 1604 1643 1781 1877 1885 2044 2046 2119 2121 2164 2165 2466 2818 3105 3216 3312 3441 3576 3659 3685 3691 3717 3837 3911 4107 4122 4179 4237 4281 4312 4664 4988

Tabella 5.2 Geni selezionati dalla procedura *shrunken centroids* applicata sul *dataset* completo con variabili standardizzate. Sono stati posti in evidenza i geni e comuni al *dataset* ridotto e quelli scelti dalla procedura ma che non appartengono alla selezione del *dataset* ridotto.

Se, facendo uso del *dataset* ridotto si applica la stessa procedura con lo stesso valore di Δ, si ha la selezione della quasi totalità dei geni dovuta ad un aumento dei coefficienti d_{ik} che regolano la selezione dei centroidi. Riferendosi all'espressione che definisce tali coefficienti, si nota che, passando da un *dataset* all'altro, l'unica quantità soggetta a variazione è il termine s_0 (definito come mediana degli s_i). L'aumento dei d_{ik} non può che essere provocato da una diminuzione di s_0, ed essendo questa la mediana degli s_i, è evidente che la distribuzione di questi ultimi è mutata rispetto al *dataset* completo. Si ricorda infine che gli s_i^2 rappresentano la somma delle varianze interne ai gruppi per ciascun gene. Pertanto si può intuire che i geni che appartengono al *dataset* ridotto sono stati selezionati in modo da minimizzare[4] la varianza interna ai gruppi delle loro espressioni (o, di contro, di massimizzare quella tra i gruppi).

Se si usa un parametro appropriato anche per il *dataset* ridotto, si riesce a selezionare un gruppo di geni che consente la classificazione delle unità con un tasso d'errore nullo. Come si nota dalla Figura 5.2, diversi valori del parametro Δ forniscono una perfetta classificazione della unità. Si è scelto di fissare $\Delta = 1{,}66$. La Tabella 5.3 fornisce la lista dei geni selezionati.

È pertanto evidente che anche all'interno del *dataset* ridotto esistono dei geni che potrebbero essere eliminati senza compromettere i risultati della classificazione. Di contro, quando si usa il *dataset* completo, alcuni geni vengono selezionati dalla procedura pur non facendo parte del *dataset* ridotto.

[4] Il termine "*minimizzare*" non viene qui usato con accezione matematica

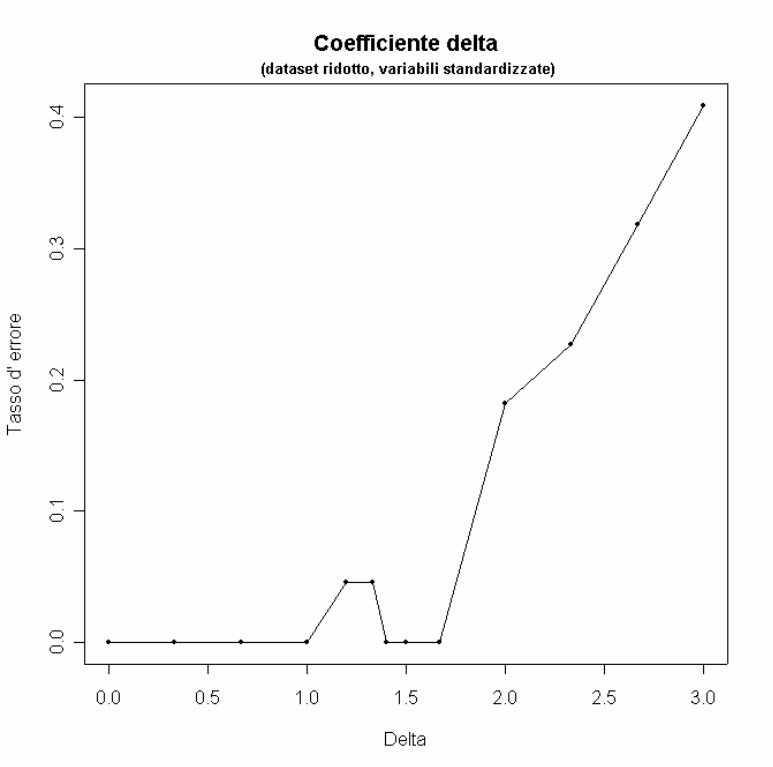

Figura 5.2 Efficacia dell'algoritmo applicato sul *dataset* ridotto con variabili standardizzate: tasso d'errore al variare del coefficiente Δ.

Variabili standardizzate, dataset ridotto		
Geni selezionati dalla procedura shrunken centroids	ALL-B	539 891 981 1180 2481 3810 4002 4060 4195 4241
	ALL-T	137 400 487 535 539 611 853 880 884 891 895 981 1180 1409 1650 2254 2441 2458 3124 3156 3253 3418 3601 3810 4002 4021 4060 4142 4178 4195 4241 4275 4420 4489 4697
	ALLTr	103 1378 1677 1695 2317 2743 3134 3257 3683 3841 3862 3987
	AML	9 452 474 558 664 674 1413 1530 1831 2103 2200 2461 2486 2848 3246 3827 3967 4110 4236 4445 4473 4508 4612 4681 4762

Tabella 5.3 Geni selezionati dalla procedura *shrunken centroid*s applicata sul *dataset* ridotto con variabili standardizzate.

In appendice si propongono le stesse analisi condotte utilizzando variabili non standardizzate (Tabella A3.1Figura A5.1e Tabella A3.1Figura A5.2)

È parso infine interessante porre a confronto i geni selezionati applicando la procedura a variabili standardizzate e non. Si propone pertanto una lista dei geni selezionati in comune sia per il *dataset* completo sia per quello ridotto (Tabella 5.4).

Geni comuni selezionati applicando l'algoritmo su variabili standardizzate e non		
Dataset completo	ALL-B	539 891 981 1180 2481 3209 3810 4002 4060 4195 4241
	ALL-T	400 487 535 539 570 611 729 853 880 891 895 981 1180 1409 1549 1650 1712 2134 2433 2697 2710 2748 2868 3124 3156 3253 3418 3601 3810 3828 4002 4021 4060 4103 4142 4178 4195 4241 4275 4420 4480 4489 4697
	ALL-Tr	57 221 601 939 1378 1553 1673 1677 1695 1730 2277 2317 2743 3134 3257 3483 3683 3841 3862 3869 3987 4112 4174
	AML	191 452 558 577 630 664 1025 1282 1530 1643 1831 1885 2016 2119 2466 2848 3312 3685 3864 3967 4107 4110 4179 4236 4237 4445 4473 4508 4525 4681 4697 4762
Dataset ridotto	ALL-B	539 891 898 981 1180 2481 3810 4002 4060 4195 4241
	ALL-T	137 400 487 535 539 611 729 853 880 884 891 895 981 1180 1409 1549 1650 1714 2254 2433 2441 2458 2697 3110 3124 3156 3253 3418 3601 3810 4002 4021 4060 4142 4178 4195 4241 4275 4420 4489 4697
	ALL-Tr	103 221 1378 1553 1677 1695 2317 2743 3134 3257 3683 3841 3862 3987
	AML	9 452 474 558 664 674 1113 1282 1413 1530 1711 1831 2019 2052 2103 2200 2461 2486 2848 2964 3246 3334 3827 3967 4110 4119 4236 4445 4473 4508 4612 4681 4762

Tabella 5.4 Geni comuni selezionati applicando l'algoritmo *shrunken centroid* su variabili standardizzate e non distini per *dataset* completo e ridotto.

Per dare un'idea visiva della selezione dei geni, si propongono i grafici delle figure 5.3 e 5.4, i primi due relativi a variabili standardizzate, gli altri a variabili non standardizzate. In entrambi i casi il grafico a sinistra riporta, i centroidi di ciascun gruppo ed in blu si evidenziano gli *"shrunken centroid"*, ossia quelli definiti come $\bar{x}_{ik}^{'}$, che, come si è detto, sono diversi da zero solo per i geni selezionati. Si noti che per variabili standardizzate, com'era da aspettarsi, i centroidi sono quasi tutti nulli mentre quelli diversi da zero corrispondono ai geni selezionati. Bisogna sottolineare però che la procedura in questo caso non seleziona tutti i geni che hanno centrodi diversi da zero, a prescindere dal loro comportamento nei quattro gruppi, seleziona piuttosto quei geni i cui centroidi hanno un comportamento diverso da gruppo a gruppo. In altre parole, se per assurdo ci fosse un gene che presenta all'incirca lo stesso valore positivo o negativo in tutti i gruppi[5], questo non verrebbe selezionato.

I grafici sulla destra si riferiscono invece ai coefficienti d_{ik} e $d_{ik}^{'}$. Si ricorda che i fattori $d_{ik}^{'}$, che vengono evidenziati in viola, sono quelli che determinano la selezione o meno del gene.

Le analisi proposte nel presente paragrafo sono state implementate nelle *routines* A5.2.r1-A5.2.r7

[5] In realtà questo non può accadere perché si sta trattando con variabili media nulla

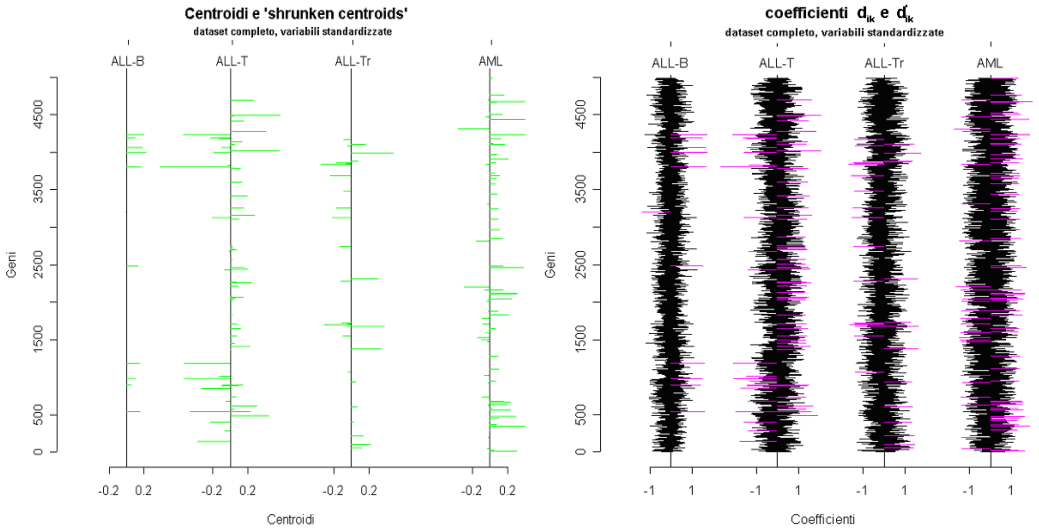

Figura 5.3 Variabili standardizzate. A sinistra: centroidi e, in verde "*shrunken centroids*" dei geni selezionati distinti per gruppo. A destra: coefficienti d_{ik} e, in viola d'_{ik} che determinano la selezione o meno di un gene.

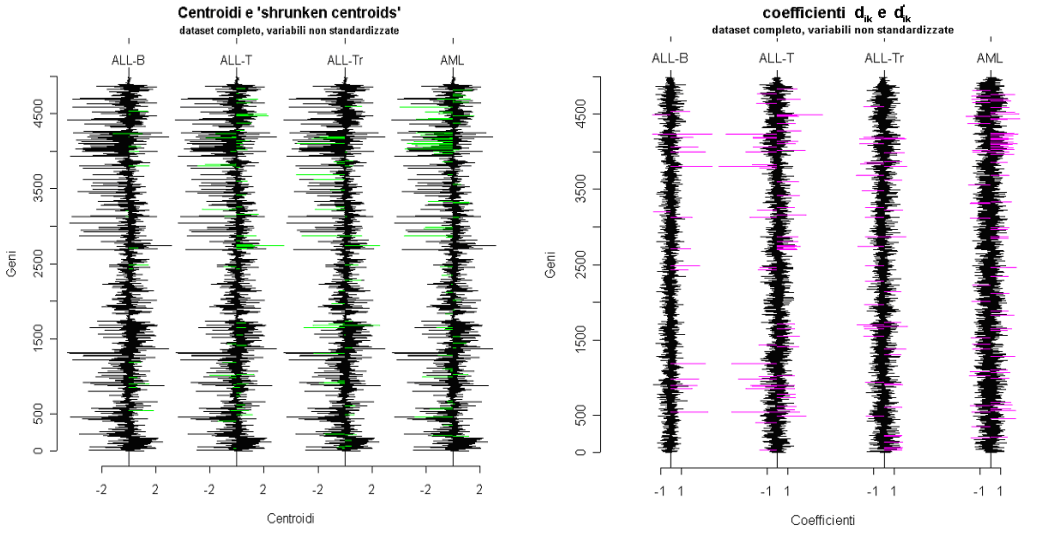

Figura 5.4 Variabili non standardizzate. A sinistra: centroidi e, in verde "shrunken cemtroids" dei geni selezionati distinti per gruppo. A destra: coefficienti d_{ik} e, in viola d'_{ik} che determinano la selezione o meno di un gene.

5.3 Confronto tra due metodi di selezione delle variabili

Nelle precedenti analisi si è constatato che, dal punto di vista statistico, una corretta classificazione non solo è possibile, ma è anche realizzabile con l'uso di un numero ridotto di variabili (geni). Dal punto di vista biologico però, si è

avuto modo di sottolineare l'importanza dell'individuazione di quei geni che, presumibilmente, sono responsabili dell'insorgere della patologia oppure, come nel caso in esame, della sua differenziazione in diverse forme.

In letteratura sono state suggerite diverse tecniche per isolare geni sovra o sotto espressi, o geni il cui comportamento risulta sospetto. Una di queste è stata applicata anche nel paragrafo precedente, e prevede la selezione di quei geni con un valore d'espressione che si discosta dalla media di gruppo sulla base di una soglia (Δ) fissata con la validazione incrociata.

Un altro criterio di selezione, introdotto da Golub *et al.* (1999) per la classificazione in due gruppi, è stato successivamente rivisto da Doudoit *et al.* (2002) e generalizzato per la classificazione in K categorie. Esso si serve del coefficiente r_i definito come

$$r_i = \left| \frac{\sum_{k=1}^{K} n_k \left(\bar{x}_{ik} - \bar{x}_i\right)^2}{\left(n-K\right) s_i^{\,2}} \right|$$

dove n_k è la numerosità del k-esimo gruppo, $\bar{x}_{ik}$ è la media delle espressioni del gene i nel gruppo k, ossia $\bar{x}_{ik} = \sum_{j \in C_k} \frac{x_{ij}}{n_k}$ (media di gruppo), $\bar{x}_i$ è la media generale del gene i calcolata su tutti i gruppi, ossia: $\bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n}$, mentre $s_i^{\,2}$ è una media pesata delle varianze di gruppo $s_{ik}^{\,2}$ definita come segue:

$$s_i^{\,2} = \frac{1}{n-K} \sum_{k=1}^{K} (n_k - 1) s_{ik}^{\,2} \,.$$

Il metodo consiste nella selezione degli m geni con valore r_i maggiore.

E'parso interessante porre a confronto questi due metodi di selezione applicandoli congiuntamente al metodo delle k-medie e dei medoidi con una procedura iterativa che ad ogni passo eliminasse il 2% dei geni sulla base di

uno dei due metodi. Con i geni selezionati ad ogni passo, si sono quindi applicati il metodo dei medoidi e delle *k*-medie[6] e, per ciascuno, si è calcolato una sorta di tasso d'errore che misurasse la corrispondenza tra i raggruppamenti creati dai due metodi e quelli corrispondenti alla reale suddivisione dei soggetti nei quattro gruppi di patologia. Come per il resto dell'elaborato, le analisi sono state condotte sul *dataset* completo e su quello ridotto ed, in entrambi i casi, si sono utilizzate variabili standardizzate e non.

Si ricorda che il metodo di Tibshirani, prevede anzitutto la definizione dei coefficienti d_{ik}:

$$d_{ik} = \frac{\overline{x}_{ik} - \overline{x}_i}{m_k (s_i + s_0)}.$$

Successivamente questi vengono confrontati con la soglia Δ attraverso la seguente funzione:

$$d'_{ik} = sign(d_{ik})\left(|d_{ik}| - \Delta\right)_+ .$$

I geni che verranno selezionati sono pertanto quelli per cui il valore d_{ik} supera la soglia Δ o, analogamente, quelli per cui d'_{ik} risulta maggiore di zero.

Il tasso d'errore, come si è detto, è stato calcolato sulla base della corrispondenza tra i raggruppamenti creati dai due metodi la reale suddivisione delle unità nei quattro gruppi. Visto che entrambi i metodi non si servono della variabile risposta ma creano raggruppamenti sulla base della somiglianza tra le unità, la definizione di tasso di errore o di unità mal allocata può non essere univoca. Nel caso in esame si è provveduto a creare ad ogni passo una tabella di frequenza del raggruppamento reale contro quello creato dal metodo in questione. A titolo di esempio, se si presentasse un *clustering* di questo tipo

[6] Anche in questo caso, per rendere stabile la convergenza dell'algoritmo delle *k*-medie, si sono fornite in input le medie di gruppo che verranno usate come centro dei *cluster* per inizializzare la procedura.

ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-T	ALL-T	ALL-T	ALL-T	ALL-T	ALL-Tr	ALL-Tr	ALL-Tr	AML	AML	AML	AML
1	1	2	1	3	1	1	1	1	3	2	2	3	1	3	3	3	3	3	3	4	3

si andrebbe a costituire la seguente tabella di frequenza

	Gruppo 1	Gruppo 2	Gruppo 3	Gruppo 4
ALL-B	7	1	2	0
ALL-T	1	2	2	0
ALL-Tr	0	0	3	0
AML	0	0	3	1

A questo punto si selezionano le caselle all'interno della tabella in modo che, detta $x_{i_1 j_1}$ la casella relativa alla patologia i_1 e al gruppo j_1 (ossia di posizione i_1, j_1), si vadano a formare tutte le (in totale 24) combinazioni del tipo $x_{i_1 j_1}\ x_{i_2 j_2}\ x_{i_3 j_3}\ x_{i_4 j_4}$ tali che, $i_1 \neq i_2 \neq i_3 \neq i_4$ *e* $j_1 \neq j_2 \neq j_3 \neq j_4$. Ciò significa che, presa una casella x_{ij} a caso, le restanti tre da selezionare non potranno appartenere né alla riga *i*, né alla colonna *j*, questo per ogni cella selezionata.
Una volta individuate tutte queste combinazioni, si calcola, per ciascun quartetto, il totale dei valori in esse contenuti. La tabella che segue illustra questa operazione per l'esempio in esame, riportando coordinate, valori e totali per ciascuna delle 24 combinazioni.

Infine, il massimo tra i totali viene considerato come il numero di osservazioni allocate correttamente. Nell'esempio, le unità che si considerano ben classificate sono tredici, ed il tasso d'errore sarà pertanto pari a 0.409.

posizione	1,1	2,2	3,3	3,4			1,2	2,4	3,1	4,3		
valori	7	2	3	1	totale	13	1	0	0	3	totale	4
posizione	1,2	2,3	3,4	4,1			1,4	2,1	3,3	4,2		
valori	1	2	0	0	totale	3	0	1	3	0	totale	4
posizione	1,3	2,4	3,1	4,2			1,1	2,3	3,2	4,4		
valori	2	0	0	0	totale	2	7	2	0	1	totale	10
posizione	1,4	2,1	3,2	4,3			1,3	2,2	3,4	4,1		
valori	0	1	0	3	totale	4	2	2	0	0	totale	4
posizione	1,4	2,3	3,2	4,1			1,3	2,1	3,4	4,2		
valori	0	2	0	0	totale	2	2	1	0	0	totale	3
posizione	1,3	2,2	3,1	4,4			1,1	2,4	3,2	4,3		
valori	2	2	0	1	totale	5	7	0	0	3	totale	10
posizione	1,2	2,1	3,4	4,3			1,4	2,2	3,3	4,1		
valori	1	1	0	3	totale	5	0	2	3	0	totale	5
posizione	1,1	2,4	3,3	4,2			1,2	2,3	3,1	4,4		
valori	7	0	3	0	totale	10	1	2	0	1	totale	4
posizione	1,4	2,3	3,1	4,2			1,2	2,1	3,3	4,4		
valori	0	2	0	0	totale	2	1	1	3	1	totale	6
posizione	1,3	2,1	3,2	4,4			1,1	2,3	3,4	4,2		
valori	2	1	0	1	totale	4	7	2	0	0	totale	9
posizione	1,1	2,2	3,4	4,3			1,3	2,4	3,2	4,1		
valori	7	2	0	3	totale	12	2	0	0	0	totale	2
posizione	1,2	2,4	3,3	4,1			1,4	2,2	3,1	4,3		
valori	1	0	3	0	totale	4	0	2	0	3	totale	5

Questo metodo consente di considerare i raggruppamenti creati dai metodi di analisi *cluster* (senza variabile risposta) nel modo più conveniente dal punto di vista del tasso d'errore, ossia consente di ricercare il raggruppamento che più si avvicina a quello reale. Naturalmente questa ricerca non è difficile nel caso di una perfetta allocazione, ma lo è quando alcune unità sono suddivise in modo dissimile dal reale raggruppamento. Ad esempio, di fronte ad un risultato come quello presentato nella tabella seguente, potrebbero sorgere dubbi riguardo alle unità contrassegnate dal numero 3 (o appartenenti al gruppo n°3):

ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-B	ALL-T	ALL-T	ALL-T	ALL-T	ALL-T	ALL-Tr	ALL-Tr	ALL-Tr	AML	AML	AML	AML
1	1	2	1	3	1	1	1	1	3	**3**	2	**3**	1	**3**	2	2	2	**3**	**3**	4	**3**

si potrebbero infatti considerare allocate correttamente tanto quelle appartenenti al gruppo AML quanto quelle appartenenti al gruppo ALL-T. Il metodo descritto sopra consente appunto di scegliere, tra le due, la combinazione che consente di ottenere il tasso d'errore minimo.[7]

I risultati di queste applicazioni vengono riassunti negli otto grafici delle figure 5.5 e 5.6. I primi quattro riguardano il metodo di Dudoit (ossia la selezione dei geni mediante i coefficienti r_i), gli altri il metodo di Tibshirani (selezione con i coefficienti d'_{ik}). Sull'asse della ascisse si riportano le percentuali di geni eliminatati, su quello delle ordinate il tasso di errata classificazione. Le analisi sopraccitate sono state realizzare attraverso le *routines* A5.2.r8-A5.2.r10-A5.2.r11-A5.2.r12.

[7] Si noti che possono esistere più combinazioni che sono caratterizzate dallo stesso tasso d'errore, in altre parole la soluzione può non essere unica.

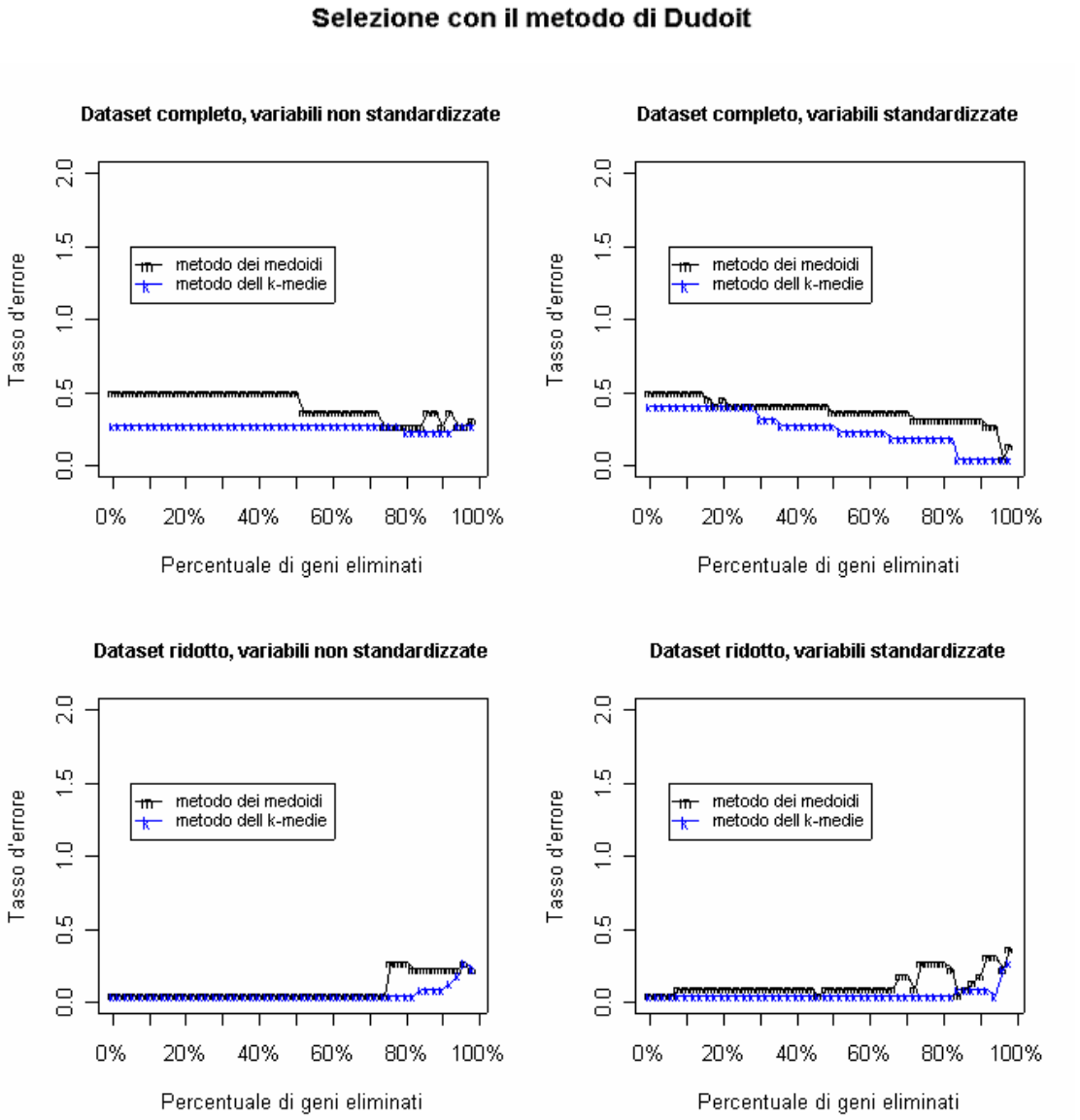

Figura 5.5 Risultati dell'applicazione dei metodi delle k-medie e dei medoidi associati alla selezione dei geni con il metodo di Dudoit.

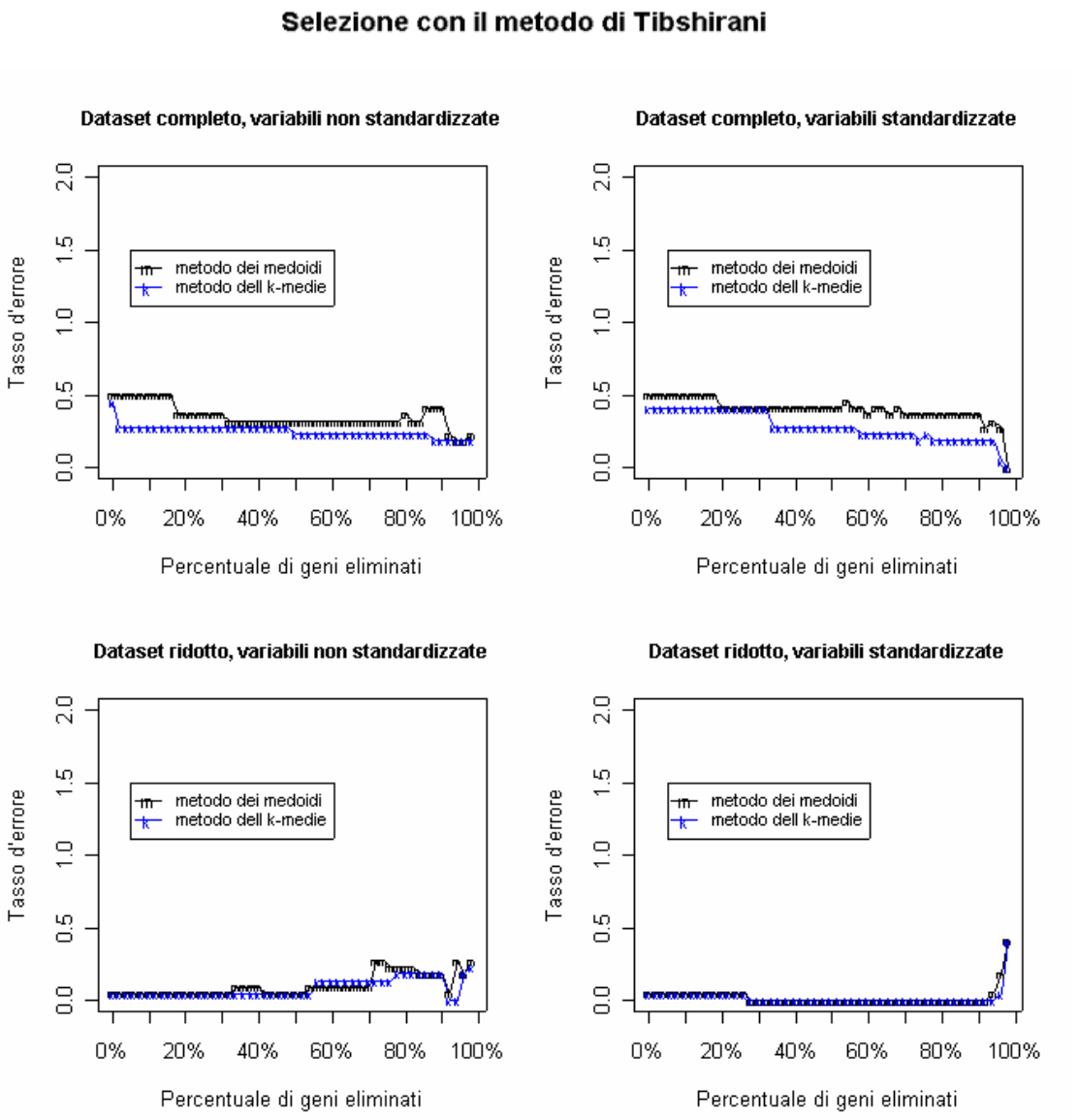

Figura 5.6 Risultati dell'applicazione dei metodi delle k-medie e dei medoidi associati alla selezione dei geni con il metodo di Tibshirani.

La prima cosa che si può notare è che nessuno dei due criteri di selezione è in grado di individuare, partendo dal *dataset* completo, un gruppo geni che, utilizzati con le due procedure di *clustering*, consentono una perfetta classificazione. Se si fa invece riferimento al *dataset* ridotto, si può constatare che il metodo di Tibshirani, consente, se applicato a variabili standardizzate, di

individuare diverse percentuali di geni che garantiscono una classificazione con tasso d'errore nullo. Anche il metodo di Dudoit, comunque, offre dei buoni risultati (tasso d'errore minimo 0.045, ossia un solo soggetto associato al gruppo sbagliato), soprattutto se applicato al *dataset* ridotto con variabili non standardizzate. Si noti infine che, soprattutto con riferimento al *dataset* completo, la classificazione trae sempre giovamento dalla progressiva eliminazione di variabili, che evidentemente, non provoca perdita di informazione ma solo riduzione del rumore.

Questo tipo di confronto, non consente comunque di deporre a favore né dell'uno né dell'altro metodo, in quanto la performance di entrambi è influenzata dall'associazione ai metodi delle k-medie e dei medoidi. Si può invece affermare che il metodo di selezione ideato da Tibshirani, se applicato al *dataset* ridotto con variabili standardizzate, è in grado di selezionare un gruppo di geni che sia con il metodo delle k-medie che con quello dei medoidi, consente, una perfetta classificazione anche con l'uso del solo 10% dei geni, mentre il metodo di selezione di Dudoit associato alle stesse procedure e nelle stesse condizioni non è in grado di selezionare delle variabili che garantiscano una perfetta classificazione. Ciò significa che, non solo all'interno del *dataset* completo ma anche nel *dataset* ridotto esistono geni che non sono essenziali per ottenere una corretta classificazione delle unità.

Va sottolineato infine che entrambi i criteri di selezione assumono implicitamente ortogonalità tra le variabili in quanto ciascun r_i (nel caso del metodo di Dudoit) e ciascun d_{ik} (metodo di Tibshirani) sono calcolati solo sulla base dei livelli assunti dal singolo gene e non tengono conto di una possibile forma di correlazione tra i geni.

5.4 Regressione logistica penalizzata

Si è detto che uno degli scopi dell'analisi delle espressioni geniche è la classificazione delle unità statistiche in classi corrispondenti a diverse forme di malattia o più semplicemente nelle categorie affetto/non affetto dalla patologia in questione. Quando le classi sono soltanto due, la regressione logistica può fornire un valido strumento per la modellazione della probabilità di appartenenza ad una classe attraverso combinazioni lineari di variabili esplicative. Purtroppo però la regressione logistica pone problemi se applicata a dati derivanti da *microarray* a causa dell'alta dimensionalità che caratterizza questi *datasets*. Come si è già avuto modo di discutere in precedenza, la presenza di un numero così elevato di variabili rispetto alle unità statistiche, fa sorgere essenzialmente due problemi: la multicollinearità e la sovraparametrizzazione. Dal primo scaturiscono sistemi di equazioni che non hanno soluzioni uniche e stabili, dal secondo in genere consegue che un modello si adatta molto bene al *dataset* rispetto al quale è stimato ma non funziona altrettanto bene se applicato per la classificazione di nuove unità. Il problema della multicollinearità nel contesto dei *microarray* è facilmente intuibile. Si è gia visto che esistono due tipi di multicollinearità: quella fondamentale e quella accidentale. La prima ha luogo perché, avendo a disposizione un così elevato numero di geni, cresce la possibilità che alcuni di essi presentino all'incirca gli stessi *pattern* di livelli alti e bassi. Pertanto è molto probabile che introducendo nuove variabili nel modello non si abbia un apporto di informazione, ma piuttosto l'insorgere di multicollinearità. La multicollinearità accidentale si può incontrare a causa della spesso limitata precisione delle misure. Molte sono le fonti di errore che si possono incontrare nei processi che portano dal colore al dato numerico. Per quanto concerne la sovraparametrizzazione, il fatto di disporre di così tante variabili rispetto alle osservazioni consente di ottenere modellazioni perfette ma purtroppo prive di senso, in quanto esistono infiniti modi di scegliere un insieme di variabili

esplicative che forniscono un perfetto adattamento del modello ma ciò non garantisce una buona prestazione previsiva su nuove unità statistiche.

Come soluzione ad entrambi i problemi, la letteratura propone la *regressione logistica con verosimiglianza penalizzata.* In questo caso non si opera una selezione di variabili, ma tutti i geni hanno lo stesso peso all'interno del modello. Ne consegue un sistema di tante equazioni quante sono le variabili. Il parametro di penalizzazione viene scelto sulla base del criterio di Akaike che sostanzialmente bilancia la capacità predittiva con la parsimonia del modello.

Dalla *log-verosimiglianza* viene pertanto sottratto un parametro che penalizza coefficienti di regressione troppo elevati a meno che essi non contribuiscano realmente a migliorare la bontà del modello. Sia y_i la variabile che stabilisce l'appartenenza o meno ad una della due classi definita come:

$$y_i = \begin{cases} 1 \ se \ i \in Classe \ 1 \\ 0 \ se \ i \in Classe \ 2 \end{cases}$$

sia inoltre

$$\Pr[y_i = 1] = p_i \quad e \quad \Pr[y_i = 0] = 1 - p_i$$

Ciò che si desidera è modellare la probabilità p_i che un individuo i con livelli di espressione rappresentati nel vettore x_i appartenga, ad esempio, alla Classe 1. Naturalmente questa probabilità non può essere stimata con un modello di regressione semplice, in quanto non si riesce a garantire che la stima appartenga all'intervallo [0, 1].

La regressione logistica si propone pertanto di modellare non p_i ma una sua trasformata h_i definita come:

$$h_i = log \frac{p_i}{1 - p_i}$$

Attraverso il seguente modello

$$h_i = log\frac{p_i}{1-p_i} = a + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip}$$

Per capire il funzionamento della penalizzazione è utile riferirsi al modello di regressione classico

$$y = X\,b + e$$

in cui i coefficienti vengono stimati con $\hat{b} = (X^T X)^{-1} X^T y$, quantità che minimizza la somma dei quadrati degli scarti

$$S = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij} b_j)^2 \ .$$

Quando esistono problemi di multicollinearità, il sistema di equazioni normali non ha una soluzione unica e, di conseguenza, alcuni o a volte tutti i coefficienti di regressione espressi nel vettore ß risultano molto elevati. La regressione penalizzata prevede di aggiungere alla forma quadratica da minimizzare, S, un termine, detto appunto di penalizzazione che ha lo scopo di scoraggiare la scelta di valori elevati dei coefficienti $ß_j$.

Hoerl e Kennard (1970) rif [57] propongono di utilizzare come penalità la somma dei quadrati dei coefficienti di regressione[8] pesati con il parametro di penalizzazione ? :

$$S^* = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij} b_j)^2 + l \sum_{j=1}^{p} b_j^2$$

Il termine $l \sum_{j=1}^{p} b_j^2$ prende il nome di *penalità.* La stima di ß diventa pertanto

$$\hat{b} = (X^T X + ?I)^{-1} X^T y \ .$$

Per quanto riguarda la regressione logistica, si è visto che il modello assume la forma

$$h_i = \log\frac{p_i}{1-p_i} = a + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip}$$

[8] Esistono anche altre scelte del termine di penalizzazione

dove p_i è la probabilità di osservare $y_i = 1$ ed h_i prende il nome di *predittore lineare.* Essendo y_i distribuita come un Bernoulli di parametro p_i , la *log-verosimiglianza* associata a questa variabile è data da:

$$\ell = \sum_{i=1}^{n} y_i \log p_i + \sum_{i=1}^{n} (1 - y_i) \log(1 - p_i)$$

mentre una *log-verosimiglianza* penalizzata può assumere questa forma

$$\ell^* = \ell - l \sum_{j=1}^{p} b_j^2 / 2 .$$

In questo caso il termine di penalizzazione $l \sum_{j=1}^{p} b_j^2 / 2$ è detto *ridge penalty*. Il parametro ? regola il peso del termine di penalizzazione: più elevato è il valore di ?, maggiore è l'influenza esercitata dal parametro di penalizzazione e più piccoli risulteranno gli elementi di *ß*. La divisione per due è solo un fatto di comodità, il termine verrà semplificato derivando la *log-verosimiglianza*.

Differenziando ℓ^* prima in a e poi in *ß* e ponendo le derivate uguali a zero, si ottiene il seguente sistema

$$\begin{cases} u^T (y - p) = 0 \\ X^T (y - p) = l\, b \end{cases}$$

in cui *u* è un vettore *p*-dimensionale di uno. Queste equazioni sono non lineari a causa della relazione non lineare tra *p* ed h. Con l'espansione fino al primo ordine attraverso la formula di Taylor si ottiene:

$$p_i \approx \tilde{p}_i + \frac{\partial p_i}{\partial a}(a - \tilde{a}) + \sum_{j=1}^{p} \frac{\partial p_i}{\partial b}(b_j - \tilde{b}_j)$$

dove le variabili contrassegnate con la tilde (˜) indicano una soluzione approssimata delle equazioni della verosimiglianza penalizzata.
Considerando che

$$\frac{\partial p_i}{\partial a} = p_i (1 - p_i)$$
$$\frac{\partial p_i}{\partial b} = p_i (1 - p_i)\, x_{ij}$$

e definendo $w_i = p_i(1-p_i)$ e $\mathbf{W} = diag(w_i)$ si ottiene

$$\begin{aligned} u'Wua + u'\tilde{W}Xb &= u'(y-\tilde{p}-\tilde{W}\tilde{h}) \\ X'Wua + (X'WX + lI)b &= X'(y-p-\tilde{W}\tilde{h}) \end{aligned} .$$

Applicando metodi iterativi si arriva ad una soluzione in modo generalmente abbastanza veloce. I valori iniziali per $\tilde{a}$ e $\tilde{b}$ generalmente scelti come $\tilde{a} = \log\left[\frac{\bar{y}}{1-\bar{y}}\right]$ e $\tilde{b} = 0$.

Si noti che il coefficiente a è strettamente legato alla frazione di soggetti che presenta la caratteristica $y_i = 1$, infatti dalla formula $\partial\ell^* / \partial a = 0$ segue che

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p_i \quad \text{da cui} \quad \bar{y} = \bar{p}$$

Il parametro di penalizzazione ? ha un'importanza fondamentale e viene scelto essere quel valore di ? che minimizza il criterio di Akaike. Tale criterio stima la bonà predittiva di un modello sottraendo il numero di parametri dalla log-verosimiglianza massimizzata, trovando così un giusto equilibrio tra la complessità del modello e la sua fedeltà ai dati.

Il criterio AIC (Akaike Information Criterion) viene pertanto definito come

$$\text{AIC} = Dev(y \mid \hat{p}) + 2Dim$$

dove $Dev(\cdot)$ è pari a -2ℓ e *Dim* è l'effettiva dimensione del modello. Hastie e Thibshirani propongono di stimare *Dim* con

$$Dim = traccia(Z(Z^T WZ + lR)^{-1} WZ^T),$$

dove $= [u \mid X]$ mentre *R* è una matrice identica $p+1 \times p+1$ con l'elemento r_{11} uguale a zero.

Per l'algoritmo di calcolo si è preso spunto da quello proposto nel sito `www.bioconductor.org` che da qualche anno si occupa di raccogliere ed

organizzare dati, articoli, package compatibili con il *software* R nonché programmare convegni in materia di elaborazione di dati da *microarray*.

Dal momento che la *routine* è stata scritta per la classificazione delle unità in soli due gruppi, si è scelto di discriminare tra leucemie ALL e AML, operazione spesso fatta anche in letteratura come passo preliminare, e poi di distinguere i due gruppi per cui si dispone di più osservazioni ossia il gruppo ALL-B ed ALL-T. Anche in questo caso, per problemi di elaborazione, non è stato possibile applicare l'algoritmo sul *dataset* completo e si è quindi fatto uso di quello ridotto.

5.4.1 Gruppo AML *vs* gruppo ALL

Per prima cosa è necessario stimare un valore conveniente del parametro di penalizzazione l. Per ottenere dei buoni risultati è stato sufficiente costituire un *training set* di sole sei unità, tre appartenenti alla classe ALL (in particolare una ALL-B, una ALL-T ed una ALL-Tr) e due alla classe AML. Per cercare un valore ottimale, l è stato fatto variare tra 1 e 10^5 usando venticinque valori equispaziati del logaritmo in base 10 di l. La Figura 5.7 illustra come varia l'indice AIC al variare del $log_{10}(l)$. Il valore ottimo per l sarà all'incirca pari a 300.

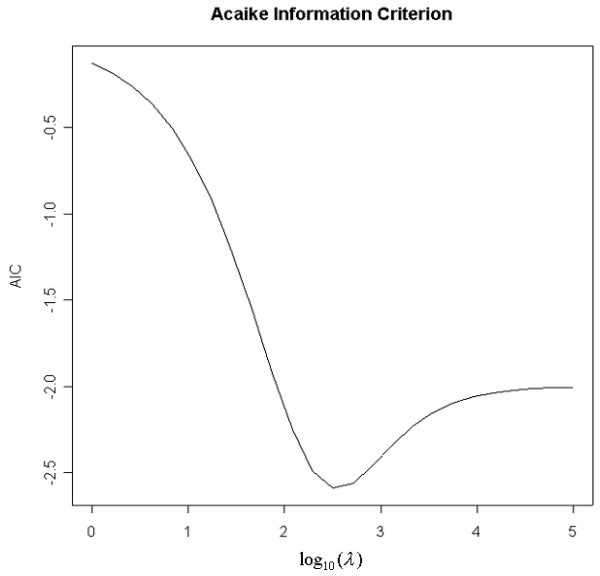

Figura 5.7 Indice AIC al variare del coefficiente di penalizzazione

Una volta fissato il parametro di penalizzazione a 300, nella fase successiva si è trattato di stimare il modello su una parte del *dataset* (*training set*) e di vagliarlo sulla rimanente (*test set*). I risultati (Tabella 5.5) sorprendono prima di tutto perché la regola sbaglia soltanto una volta sia nel *test* che nel *training set* e poi perché nonostante le dimensioni ridotte del *training set* la regola di allocazione che su di esso viene costruita è in grado di dare dei risultati molto buoni.

Regressione logistica penalizzata							
Training set		Classe reale		Test set		Classe reale	
		ALL	AML			ALL	AML
Classe prevista	ALL	3	1	Classe prevista	ALL	15	1
	AML	0	1		AML	0	1
Tasso d'errore:		0,2		Tasso d'errore:		0,059	

Tabella 5.5 Matrice di confusione relativa all'applicazione della regressione logistica penalizzata, la pare a sinistra si riferisce al training set, quella a destra al test set

La Figura 5.8 fornisce una visione dei coefficienti stimati. Il grafico a destra si riferisce alla quantità $\hat{b}_i \hat{s}_i$ dove $\hat{s}_i$ è la stima della deviazione standard del gene *i* . Infatti, dal momento che il contributo del predittore lineare h_i è dato da $\hat{b}_i x_i$, allora $\hat{b}_i \hat{s}_i$ è un buon indicatore dell'influenza della colonna (ossia del gene) *i.*

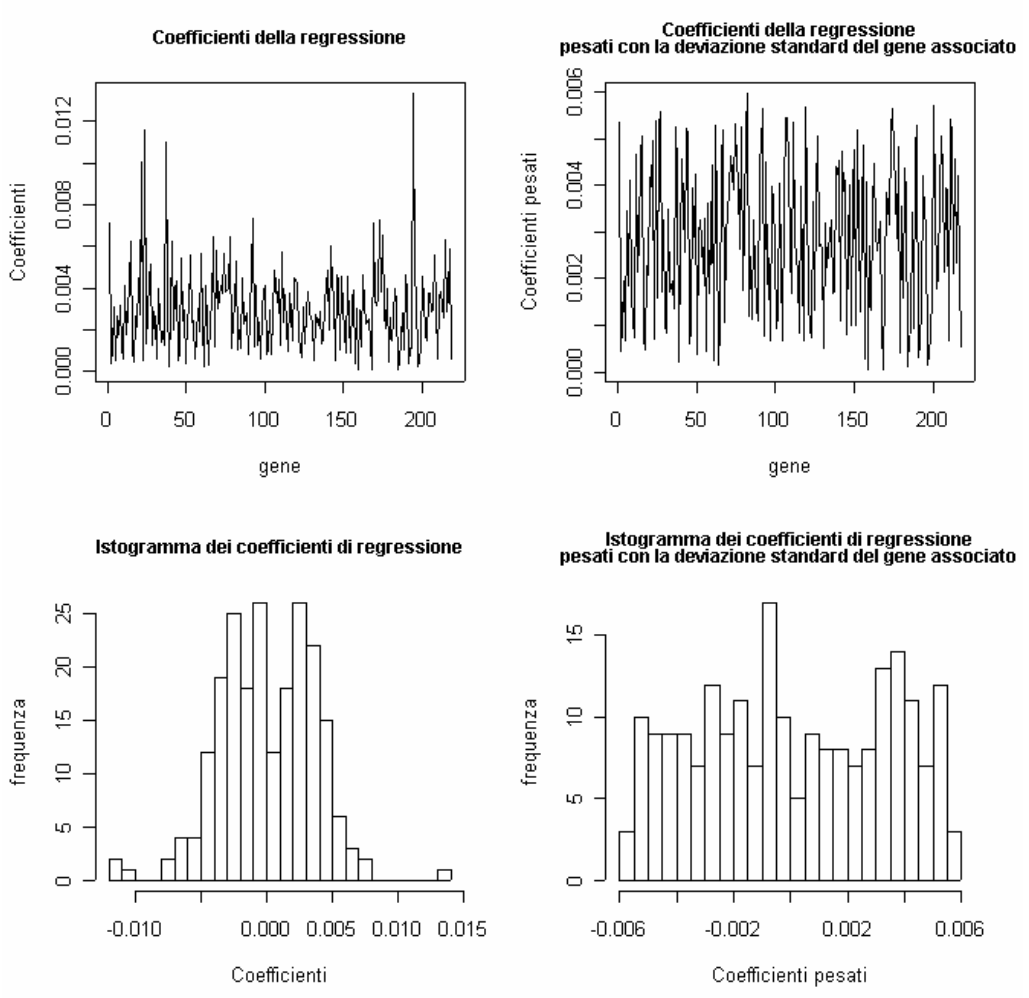

Figura 5.8 Rappresentazione dei coefficienti stimati e relativi istogrammi. La parte a destra si riferisce alla quantità $\hat{b}_i \hat{s}_i$ dove $\hat{s}_i$ è la stima della deviazione standard del gene *i*.

La Figura 5.9 riporta invece la stima della probabilità che un soggetto presenti la leucemia di tipo AML contro la variabile risposta y definita come

$$y_i = \begin{cases} 1 & se\ i \in classe\ AML \\ 0 & se\ i \in classe\ ALL \end{cases}$$

I punti colorati in rosso corrispondono alle unità che costituiscono il *learning set*, quelli colorati in blu rappresentano il *training set*. Sono stati usati dei piccoli spostamenti verticali per evitare la sovrapposizione dei punti. La linea verticale tratteggiata è stata tracciata in corrispondenza di una probabilità pari a 0,5, soglia che discrimina l'appartenenza all'una o all'altra classe.

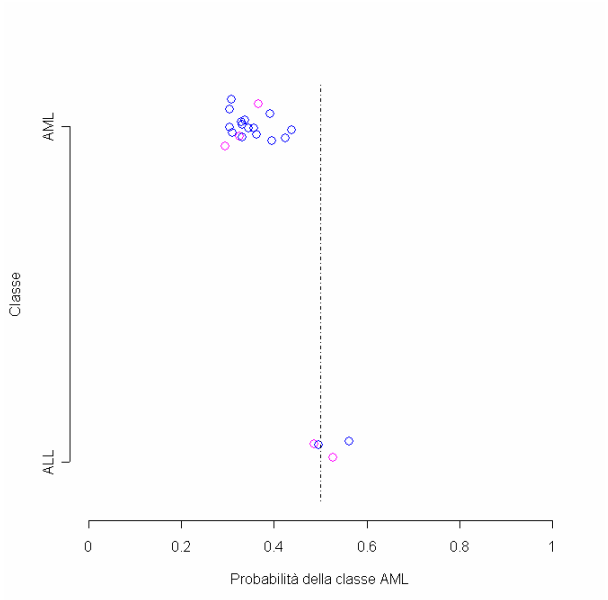

Figura 5.9 Stima della probabilità che un individuo appartenga alla classe AML

5.4.2 Gruppo ALL-B vs gruppo ALL-T

Per ottenere dei buoni risultati è stato sufficiente costituire un *training set* di sole sei unità, tre appartenenti alla classe ALL-B e tre alla classe ALL-T. l è

stato fatto variare ancora tra 1 e 10^5 usando venticinque valori equispaziati del logaritmo in base 10 di l. La Figura 5.10 illustra come l'indice AIC al variare del $\log_{10}(l)$. Il valore ottimo per l questa volta è all'incirca pari a 500.

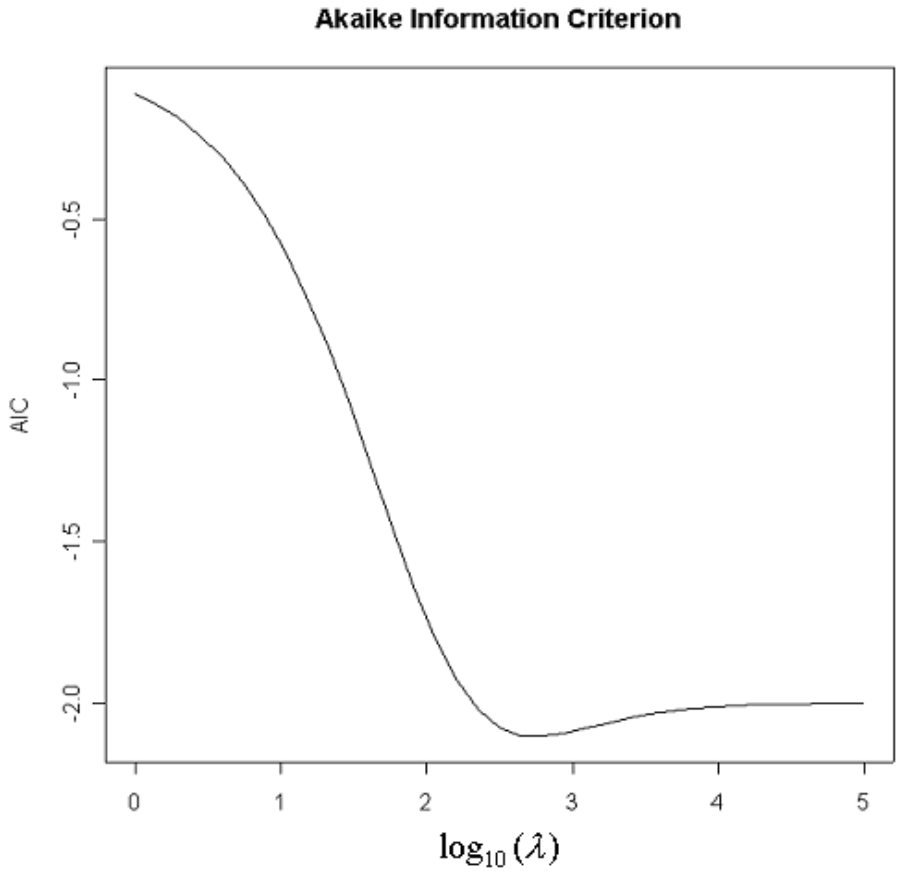

Figura 5.10 Indice AIC al variare del coefficiente di penalizzazione

Nuovamente si è stimato il modello sul *training set* e lo si è testato sul *test set*. I risultati (Tabella 5.6) sorprendono prima di tutto perché la regola ha tasso d'errore nullo sia nel *test* che nel *training set* e poi perché nonostante le dimensioni ridotte del *training set* la regola di allocazione che su di esso viene costruita è in grado di dare dei risultati molto buoni.

È tuttavia necessario osservare che sia in questo caso che nel precedente (ALL *vs* AML) la buona capacità predittiva del modello può dipendere anche

dal fatto che il *dataset* è costituito da poche osservazioni e dunque è molto facile che una funzione interpoli bene i dati.

Regressione logistica penalizzata							
Training set		Classe reale		Test set		Classe reale	
		ALL-B	ALL-T			ALL-B	ALL-T
Classe prevista	ALL-B	3	0	Classe prevista	ALL-B	7	0
	ALL-T	0	3		ALL-T	0	2
Tasso d'errore: 0				Tasso d'errore: 0			

Tabella 5.6 Matrice di confusione relativa all'applicazione della regressione logistica penalizzata, la pare a sinistra si riferisce al training set, quella a destra al test set

La Figura 5.11 fornisce una visione dei coefficienti stimati. Il grafico a destra si riferisce alla quantità $\hat{b}_i \hat{s}_i$ dove $\hat{s}_i$ è la stima della deviazione standard del gene *i* . Infatti, dal momento che il contributo del predittore lineare h_i è dato da $\hat{b}_i x_i$, allora $\hat{b}_i \hat{s}_i$ è un buon indicatore dell'influenza della colonna (ossia del gene) *i.*

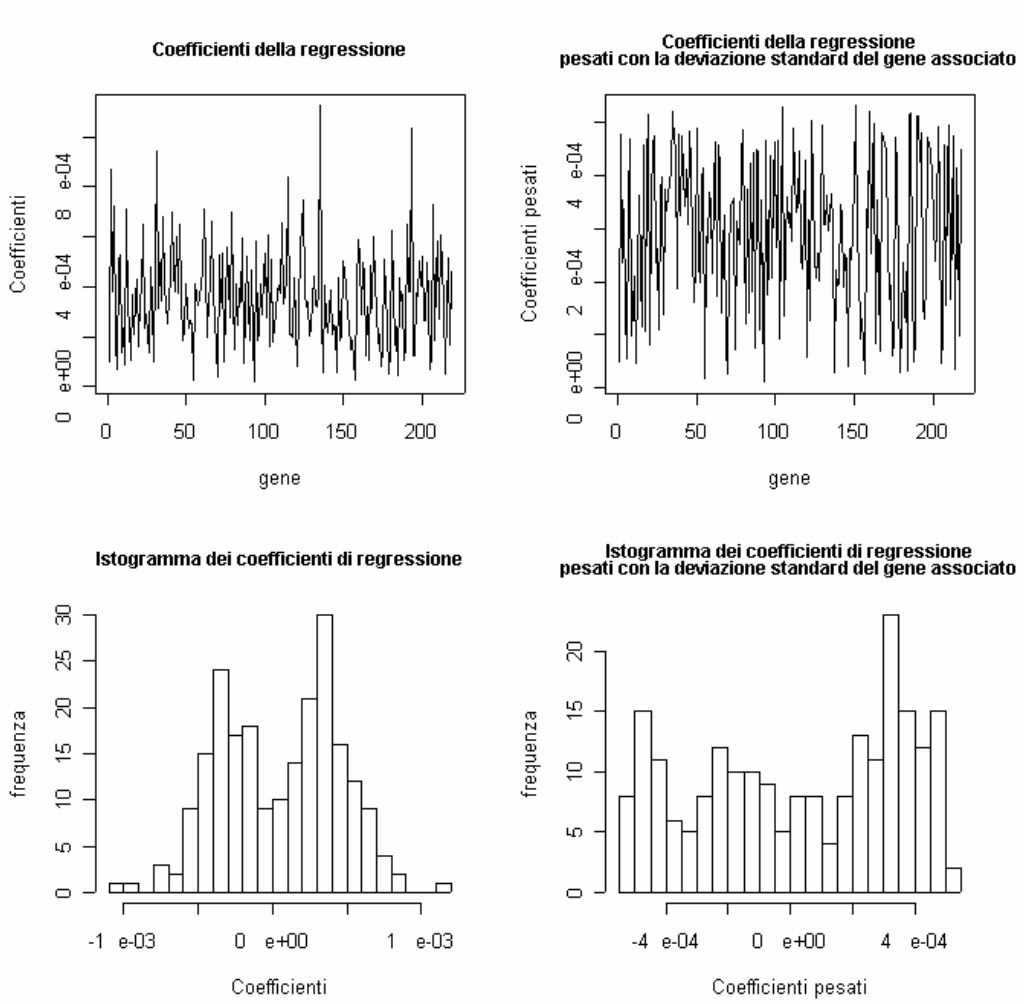

Figura 5.11 Rappresentazione dei coefficienti stimati e relativi istogrammi. La parte a destra si riferisce alla quantità $\hat{b}_i \hat{s}_i$ dove $\hat{s}_i$ è la stima della deviazione standard del gene *i*.

La Figura 5.12 riporta la stima della probabilità che un soggetto presenti la leucemia di tipo AML contro la variabile risposta y definita come

$$y_i = \begin{cases} 1 & se\ i \in classe\ ALL-B \\ 0 & se\ i \in classe ALL-T \end{cases}$$

come nel caso precedente, i punti colorati in rosso corrispondono alle unità che costituiscono il *learning set*, quelli colorati in blu rappresentano il *training set*.

Piccoli spostamenti verticali garantiscono che i punti non si sovrappangano. La linea verticale tratteggiata, è stata tracciata in corrispondenza di una probabilità pari a 0,5, soglia che discrimina l'appartenenza all'una o all'altra classe.

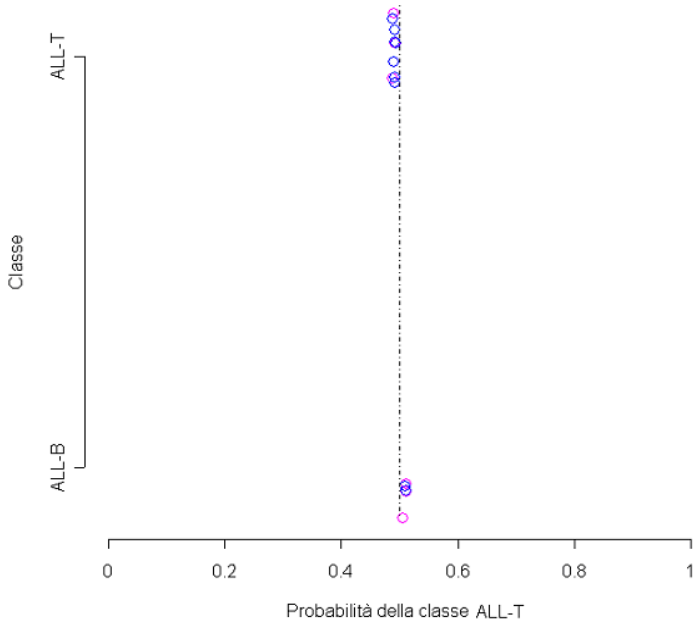

Figura 5.12 Stima della probabilità che un individuo appartenga alla classe ALL-T

Appendice

A5.1 Metodo "*shrunken centroids*"

Risultati dell'analisi *shrunken centroid*s condotta su variabili non standardizzate e sul *dataset* completo e ridotto.

Per il *dataset* completo si è fissato $\Delta = 1{,}66$ per quello ridotto $\Delta = 1$; entrambi sono stati scelti sulla base delle Figura A5.1 e Figura A5.2. Le Tabella A.5.1Tabella A.5.2 forniscono le liste dei geni selezionati dalla procedura nonché i confronti con i geni selezionati dai biologi per il *dataset* ridotto.

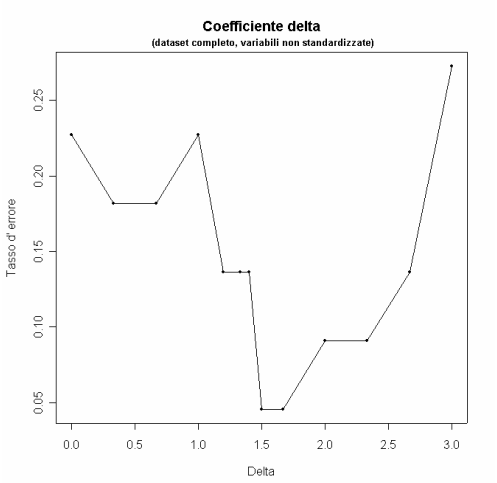

Figura A5.1 Efficacia dell'algoritmo *shrunken centroid* applicato sulle variabili non standardizzate del *dataset* completo al variare del coefficiente Δ.

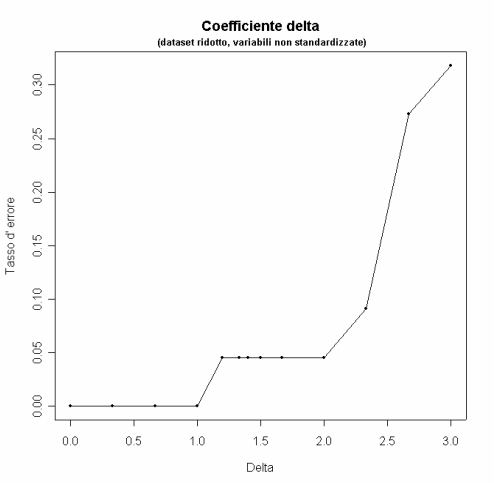

Figura A5.2 Efficacia dell'algoritmo *shrunken centroid* applicato sulle variabili non standardizzate del *dataset* ridotto al variare del coefficiente Δ.

Variabili non standardizzate, dataset completo																
Geni selezionati dalla procedura shrunken centroids	ALL-B	539	853	891	898	981	1180	2433	2481	2714	3124	3209	3810	4002	4060	4195
		4236	4241	4489	4535											
	ALL-T	29	400	487	535	539	570	611	729	771	852	853	880	891	895	981
		1012	1025	1180	1409	1429	1549	1650	1712	2433	2465	2697	2710	2735	2743	2748
		2804	2868	3124	3156	3224	3253	3418	3601	3780	3810	3828	4002	4021	4060	4065
		4103	4142	4177	4178	4195	4241	4275	4380	4473	4480	4489	4646	4697	4769	4840
	ALLTr	29	57	67	142	221	228	487	601	900	915	936	1300	1378	1549	1553
		1645	1673	1677	1695	1730	1964	2141	2277	2481	2743	2866	2868	3134	3224	3257
		3483	3619	3683	3810	3841	3862	3869	3873	3987	4105	4168	4174	4180	4189	4601
	AML	191	205	349	452	536	558	577	628	630	664	990	1025	1065	1071	1093
		1282	1376	1437	1530	1620	1643	1831	1885	2016	2119	2228	2344	2466	2848	2866
		2965	2984	3117	3312	3326	3334	3563	3685	3864	3967	4002	4003	4021	4029	4030
		4049	4076	4094	4097	4100	4102	4105	4107	4110	4133	4162	4167	4168	4179	4215
		4236	4237	4245	4380	4428	4445	4473	4508	4525	4589	4653	4681	4697	4698	4745
		4762	4820													
Geni selezionati e comuni al dataset ridotto	ALL-B	539	853	891	898	981	1180	2433	2481	2714	3124	3209	3810	4002	4060	4195
		4236	4241	4489												
	ALL-T	29	400	487	535	539	570	611	729	771	853	880	891	895	981	1012
		1025	1180	1409	1429	1549	1650	1712	2433	2697	2710	2743	2748	2804	2868	3124
		3156	3224	3253	3418	3601	3810	3828	4002	4021	4060	4065	4103	4142	4178	4195
		4241	4275	4380	4473	4480	4489	4646	4697	4769						
	ALL-Tr	29	142	221	487	900	1378	1549	1553	1677	1695	1730	1964	2481	2743	2866
		2868	3134	3224	3257	3483	3683	3810	3841	3862	3869	3987	4105	4168	4174	
	AML	452	558	664	1025	1282	1530	1831	2016	2848	2866	3326	3334	3864	3967	4002
		4021	4029	4049	4100	4105	4110	4162	4167	4168	4215	4236	4380	4445	4473	4508
		4525	4681	4697	4762	4820										
geni selezionati e non comuni al dataset ridotto	ALL-B	4535														
	ALL-T	35	3780	4177	4840											
	ALL-Tr	57	67	228	601	915	936	1300	1645	1673	2141	2277	3619	3873	4180	4189
		4601														
	AML	191	205	349	536	577	628	630	990	1065	1071	1093	1376	1437	1620	1643
		1885	2119	2228	2344	2466	2965	2984	3117	3312	3563	3685	4003	4030	4076	4094
		4097	4102	4107	4133	4179	4237	4245	4428	4589	4653	4698	4745			

Tabella A.5.1 Geni selezionati dalla procedura *shrunken centroid*s applicata sul *dataset* completo con variabili non standardizzate. Sono stati posti in evidenza i geni selezionati dalla proceduta e comuni al *dataset* ridotto e quelli scelti dalla procedure ma che non appartengono alla selezione del *dataset* ridotto.

Variabili non standardizzate, dataset ridotto																	
Geni selezionati dalla procedura shrunken centroids	ALL-B	65	128	539	729	771	797	853	880	891	898	900	981	1180	1261	1282	1413
		1452	1708	1964	2433	2481	2714	2804	2927	3124	3209	3467	3683	3810	3828	4002	4029
		4060	4065	4069	4100	4119	4164	4167	4168	4178	4195	4236	4241	4407	4445	4489	4507
		4508	4525	4603	4612	4646	4673	4760	4762	4769	4891						
	ALL-T	29	134	137	287	400	487	535	539	558	570	611	661	708	720	729	730
		771	853	880	884	891	895	913	981	1012	1025	1033	1180	1244	1261	1282	1323
		1324	1364	1374	1409	1429	1447	1452	1549	1650	1712	1714	1887	2016	2034	2051	2244
		2250	2254	2345	2433	2441	2458	2481	2673	2697	2710	2714	2732	2743	2748	2804	2812
		2849	2866	2868	3110	3124	3156	3209	3224	3253	3326	3334	3418	3442	3601	3710	3806
		3810	3828	3967	4002	4014	4015	4021	4043	4049	4060	4065	4092	4103	4142	4149	
		4168	4178	4195	4215	4241	4275	4324	4333	4380	4407	4420	4473	4480	4489	4646	4162
		4769	4891														4697
	ALLTr	29	103	128	134	142	212	221	287	487	535	565	730	797	890	891	895
		898	900	913	1012	1113	1183	1282	1323	1378	1549	1553	1619	1650	1677	1695	1710
		1730	1887	1964	2207	2317	2481	2649	2743	2849	2866	2868	2927	3124	3134	3224	3257
		3326	3483	3683	3710	3810	3841	3862	3869	3987	4069	4105	4110	4130	4149	4162	4167
		4168	4174	4218	4229	4331	4333	4441	4473	4507	4508	4525	4612	4646	4820	4858	
	AML	9	21	65	128	287	452	474	506	539	558	565	621	664	674	720	890
		895	898	900	913	1012	1025	1033	1113	1183	1226	1282	1324	1374	1413	1471	1530
		1567	1650	1708	1711	1730	1831	1833	1887	2016	2019	2052	2103	2200	2207	2250	2345
		2461	2481	2484	2486	2517	2649	2673	2710	2714	2732	2748	2812	2848	2866	2927	2964
		3156	3224	3246	3326	3334	3390	3442	3467	3629	3806	3810	3827	3864	3967	4002	4014
		4015	4021	4029	4049	4060	4092	4100	4105	4110	4119	4130	4145	4149	4162	4164	4167
		4168	4215	4218	4229	4236	4241	4380	4445	4473	4480	4489	4507	4508	4525	4612	4646
		4673	4681	4697	4762	4820											

Tabella A.5.2 Geni selezionati dalla procedura *shrunken centroid*s applicata sul *dataset* ridotto con variabili non standardizzate.

A5.2 *Routines* R

A5.2.r1 `#algoritmo srunken centroids`

```
sr.centroids<-function
(ga,gb,gc,gd,data,prior=c(1/4,1/4,1/4,1/4),delta,x)
{
data<-t(data)
ga<-t(ga)
gb<-t(gb)
gc<-t(gc)
gd<-t(gd)
mg<-apply(data,1,mean)
ma<-apply(ga,1,mean)
mb<-apply(gb,1,mean)
mc<-apply(gc,1,mean)
md<-apply(gd,1,mean)
ssa<-apply(((ga-ma)**2),1,sum)
ssb<-apply(((gb-mb)**2),1,sum)
ssc<-apply(((gc-mc)**2),1,sum)
ssd<-apply(((gd-md)**2),1,sum)
si<-sqrt((1/(ncol(data)-4))*(ssa+ssb+ssc+ssd))
so<-median(si)
m<-rep(0,4)
m[1]<-sqrt((1/ncol(ga))+(1/ncol(data)))
m[2]<-sqrt((1/ncol(gb))+(1/ncol(data)))
m[3]<-sqrt((1/ncol(gc))+(1/ncol(data)))
m[4]<-sqrt((1/ncol(gd))+(1/ncol(data)))
da<-(ma-mg)/(m[1]*(si+so))
db<-(mb-mg)/(m[2]*(si+so))
dc<-(mc-mg)/(m[3]*(si+so))
dd<-(md-mg)/(m[4]*(si+so))
p1<-abs(da)-delta
p2<-abs(db)-delta
p3<-abs(dc)-delta
p4<-abs(dd)-delta
dda<-parte.pos(p1)$y
ddb<-parte.pos(p2)$y
ddc<-parte.pos(p3)$y
ddd<-parte.pos(p4)$y

dda<-sign(da)*dda
ddb<-sign(db)*ddb
ddc<-sign(dc)*ddc
ddd<-sign(dd)*ddd

xa<-mg+m[1]*(si+so)*dda
xb<-mg+m[2]*(si+so)*ddb
xc<-mg+m[3]*(si+so)*ddc
xd<-mg+m[4]*(si+so)*ddd
punt<-rep(0,4)
punt[1]<-sum(((x-xa)**2)/((si+so)**2))-2*log(prior[1])
punt[2]<-sum(((x-xb)**2)/((si+so)**2))-2*log(prior[2])
punt[3]<-sum(((x-xc)**2)/((si+so)**2))-2*log(prior[3])
```

```
punt[4]<-sum(((x-xd)**2)/((si+so)**2))-2*log(prior[4])
pa<-parte.pos(abs(da)-delta)$pos
pb<-parte.pos(abs(db)-delta)$pos
pc<-parte.pos(abs(dc)-delta)$pos
pd<-parte.pos(abs(dd)-delta)$pos
gr<-which.min(punt)
list(gruppo=gr,punteggi=punt,genia=pa,genib=pb,genic=pc,
genid=pd)
}
```

A5.2.r2

```
#Funzione parte positiva

parte.pos<-function (x)
{
r<-NULL
y<-NULL
z<-0
j<-0
for(i in 1:length(x)){
z<-z+1
if(x[i]>=0){
j<-j+1
r[j]<-z
y[i]<-x[i]
}
if(x[i]<0){
y[i]<-0
}
}
list(y=y,pos=r)
}
```

A5.2.r3

```
#cross validation con l'algoritmo srunken centroids

sr.centroids.cv<-function
(ga,gb,gc,gd,data,prior=c(1/4,1/4,1/4,1/4),delta)
{
vn<-c(1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,3,3,3,4,4,4,4)
me<-0
data<-t(data)
ga<-t(ga)
gb<-t(gb)
gc<-t(gc)
gd<-t(gd)
gruppo<-NULL
for (i in 1:ncol(data)){
    me<-me+1
    x<-data[,me]
    data1<-data[,-me]
    if(me<=10){
      gga<-ga[,-me]
      ggb<-gb
      ggc<-gc
```

```
      ggd<-gd
      }
    if((me>10)&&(me<=15)){
      mme<-me-10
      ggb<-gb[,-mme]
      gga<-ga
      ggc<-gc
      ggd<-gd
      }
    if((me>15)&&(me<=18)){
      mme<-me-15
      ggc<-gc[,-mme]
      ggb<-gb
      gga<-ga
      ggd<-gd
      }
    if(me>18){
      ggd<-gd[,-mme]
      mme<-me-18
      ggb<-gb
      ggc<-gc
      gga<-ga
      }
    gruppo[i]<-
sr.centroids(t(gga),t(ggb),t(ggc),t(ggd),t(data1),prior=
prior,delta=delta,x)$gruppo
    }
g<-
sr.centroids(t(gga),t(ggb),t(ggc),t(ggd),t(data1),prior=
prior,delta=delta,x)
f<-rep(1,22)
err<-(22-sum(f[gruppo==vn]))/22
list(ragg=gruppo,tasso.errore=err,genia=g$genia,genib=g$
genib,genic=g$genic,genid=g$genid)
}
```

A5.2.r4

```
#algoritmo srunken centroids con la cross validation e
diversi valori di delta

rep.srcentr<-function
(ga,gb,gc,gd,data,prior=c(1/4,1/4,1/4,1/4),delta)
{
tasso.err<-NULL
for (i in 1:length(delta)){
tasso.err[i]<-
sr.centroids.cv(ga,gb,gc,gd,data,prior=prior,delta[i])$t
asso.errore
}
tasso.err
}


col[cl==2]<-c2
col[cl==3]<-c3
col1<-NULL
```

```
for (i in 1:length(a)){
    cola[i]<-col[a[i]]
    }
x
for (i in 1:length(b)){
    colb[i]<-col[b[i]]
    }
x
for (i in 1:length(c)){
    colc[i]<-col[c[i]]
    }
x
iniz<-rep(x0,ngen)
fine<-rep(x1,ngen)
segments(x0=iniz,x1=fine,y0=c(1:ngen),y1=c(1:ngen),col=c
(cola,colb,colc))
}
```

A5.2.r5

```
#confronta i geni selezionati con il dataset completo o
ridotto

tav.cfr<-function(xs1a,x1a,xs2a,x2a,xs3a,x3a,xs4a,x4a){
y11<-geni.sel(x1a,xs1a)
y12<-geni.sel(x1a,xs2a)
y13<-geni.sel(x1a,xs3a)
y14<-geni.sel(x1a,xs4a)

y21<-geni.sel(x2a,xs1a)
y22<-geni.sel(x2a,xs2a)
y23<-geni.sel(x2a,xs3a)
y24<-geni.sel(x2a,xs4a)

y31<-geni.sel(x3a,xs1a)
y32<-geni.sel(x3a,xs2a)
y33<-geni.sel(x3a,xs3a)
y34<-geni.sel(x3a,xs4a)

y41<-geni.sel(x4a,xs1a)
y42<-geni.sel(x4a,xs2a)
y43<-geni.sel(x4a,xs3a)
y44<-geni.sel(x4a,xs4a)

z<-matrix(rep(0,16),ncol=4)

z[1,1]<-length(y11)
z[1,2]<-length(y12)
z[1,3]<-length(y13)
z[1,4]<-length(y14)
z[2,1]<-length(y21)
z[2,2]<-length(y22)
z[2,3]<-length(y23)
z[2,4]<-length(y24)
z[3,1]<-length(y31)
z[3,2]<-length(y32)
z[3,3]<-length(y33)
z[3,4]<-length(y34)
```

```
z[4,1]<-length(y41)
z[4,2]<-length(y42)
z[4,3]<-length(y43)
z[4,4]<-length(y44)

z
}
```

A5.2.r6

```
#Dati due vettori seleziona gli elementi comuni

cfr<-function (a,b)
{
a<-as.vector(a)
b<-as.vector(b)
x<-0
y<-NULL
for (i in 1:length(a)){
    for(j in 1:length(b)){
       if(a[i]==b[j]){
         x<-x+1
         y[x]<-a[i]
         }
       }

    }
y
}
```

A5.2.r7

```
#Dati due vettori seleziona gli elementi di a non comuni
al vettore b

diversi<-function (a,b)
{
a<-as.vector(a)
b<-as.vector(b)
y<-NULL
k<-0
for (i in 1:length(a)){
    x<-0
    for(j in 1:length(b)){

       if(a[i]==b[j]){
         x<-1
         }
       }
    if(x==0){
      k<-k+1
      y[k]<-a[i]}
    }
y
}
```

A5.2.r8

```
rhoi<-function(dati,ga,gb,gc,gd){
ma<-apply(ga,2,mean)
mb<-apply(gb,2,mean)
mc<-apply(gc,2,mean)
md<-apply(gd,2,mean)
mg<-apply(dati,2,mean)
na<-nrow(ga)
nb<-nrow(gb)
nc<-nrow(gc)
nd<-nrow(gd)
n<-na+nb+nc+nd
dev1<-na*(ma-mg)^2
dev2<-nb*(mb-mg)^2
dev3<-nc*(mc-mg)^2
dev4<-nd*(md-mg)^2
sig1<-apply(ga,2,var)
sig2<-apply(gb,2,var)
sig3<-apply(gc,2,var)
sig4<-apply(gd,2,var)
sigma<-(1/(n-4))*((na-1)*sig1+(nb-1)*sig2+(nc-
1)*sig3+(nd-1)*sig4)
rho<-abs((na*((ma-mg)^2)+nb*((mb-mg)^2)+nc*((mc-
mg)^2)+nd*((md-mg)^2))/(n-4)*sigma)
rho
}
```

A5.2.r9

```
rho.k<-function(dati,ga,gb,gc,gd,prob){
a1<-c(1,2,3,4,4,3,2,1,2,4,1,3,3,1,4,2,4,3,1,2,2,1,3,4)
a2<-c(2,3,4,1,3,2,1,4,4,1,3,2,1,4,2,3,3,1,2,4,1,3,4,2)
a3<-c(3,4,1,2,2,1,4,3,1,3,2,4,4,2,3,1,1,2,4,3,3,4,2,1)
a4<-c(4,1,2,3,1,4,3,2,3,2,4,1,2,3,1,4,2,4,3,1,4,2,1,3)
perm<-rbind(a1,a2,a3,a4)

te<-rep(0,length(prob))
rho<-rhoi(dati,ga,gb,gc,gd)
for(i in 1:length(prob)){
    quant<-quantile(rho,prob=prob[i])
    sel<-c(1:ncol(dati))
    sel<-sel[rho>quant]
    data<-dati[,sel]
    m1<-apply(data[1:10,],2,mean)
    m2<-apply(data[11:15,],2,mean)
    m3<-apply(data[16:18,],2,mean)
    m4<-apply(data[19:22,],2,mean)
    med<-rbind(m1,m2,m3,m4)
    ris<-kmeans(data,med)$cluster
    t1<-rep(0,4)
    t2<-rep(0,4)
    t3<-rep(0,4)
    t4<-rep(0,4)

    t1[1]<-sum(rep(1,10)[ris[1:10]==1])
```

```
    t1[2]<-sum(rep(1,10)[ris[1:10]==2])
    t1[3]<-sum(rep(1,10)[ris[1:10]==3])
    t1[4]<-sum(rep(1,10)[ris[1:10]==4])

    t2[1]<-sum(rep(1,5)[ris[11:15]==1])
    t2[2]<-sum(rep(1,5)[ris[11:15]==2])
    t2[3]<-sum(rep(1,5)[ris[11:15]==3])
    t2[4]<-sum(rep(1,5)[ris[11:15]==4])

    t3[1]<-sum(rep(1,3)[ris[16:18]==1])
    t3[2]<-sum(rep(1,3)[ris[16:18]==2])
    t3[3]<-sum(rep(1,3)[ris[16:18]==3])
    t3[4]<-sum(rep(1,3)[ris[16:18]==4])

    t4[1]<-sum(rep(1,4)[ris[19:22]==1])
    t4[2]<-sum(rep(1,4)[ris[19:22]==2])
    t4[3]<-sum(rep(1,4)[ris[19:22]==3])
    t4[4]<-sum(rep(1,4)[ris[19:22]==4])

    t<-rbind(t1,t2,t3,t4)
    val<-NULL
    w<-matrix(0,ncol=4,nrow=24)
    for(j in 1:24){
       ind<-perm[,j]
       w[j,]<-
c(t[1,ind[1]],t[2,ind[2]],t[3,ind[3]],t[4,ind[4]])
       val[j]<-
sum(t[1,ind[1]],t[2,ind[2]],t[3,ind[3]],t[4,ind[4]])
       }
    val
    te[i]<-(22-max(val))/22

    }
te
}
```

A5.2.r10

```
rho.pam<-function(dati,ga,gb,gc,gd,prob){
a1<-c(1,2,3,4,4,3,2,1,2,4,1,3,3,1,4,2,4,3,1,2,2,1,3,4)
a2<-c(2,3,4,1,3,2,1,4,4,1,3,2,1,4,2,3,3,1,2,4,1,3,4,2)
a3<-c(3,4,1,2,2,1,4,3,1,3,2,4,4,2,3,1,1,2,4,3,3,4,2,1)
a4<-c(4,1,2,3,1,4,3,2,3,2,4,1,2,3,1,4,2,4,3,1,4,2,1,3)
perm<-rbind(a1,a2,a3,a4)

te<-rep(0,length(prob))
rho<-rhoi(dati,ga,gb,gc,gd)
for(i in 1:length(prob)){
    quant<-quantile(rho,prob=prob[i])
    sel<-c(1:ncol(dati))
    sel<-sel[rho>quant]
    data<-dati[,sel]
    ris<-pam(data,4,diss=F)$clustering
```

```
    t1<-rep(0,4)
    t2<-rep(0,4)
    t3<-rep(0,4)
    t4<-rep(0,4)
    t1[1]<-sum(rep(1,10)[ris[1:10]==1])
    t1[2]<-sum(rep(1,10)[ris[1:10]==2])
    t1[3]<-sum(rep(1,10)[ris[1:10]==3])
    t1[4]<-sum(rep(1,10)[ris[1:10]==4])

    t2[1]<-sum(rep(1,5)[ris[11:15]==1])
    t2[2]<-sum(rep(1,5)[ris[11:15]==2])
    t2[3]<-sum(rep(1,5)[ris[11:15]==3])
    t2[4]<-sum(rep(1,5)[ris[11:15]==4])

    t3[1]<-sum(rep(1,3)[ris[16:18]==1])
    t3[2]<-sum(rep(1,3)[ris[16:18]==2])
    t3[3]<-sum(rep(1,3)[ris[16:18]==3])
    t3[4]<-sum(rep(1,3)[ris[16:18]==4])

    t4[1]<-sum(rep(1,4)[ris[19:22]==1])
    t4[2]<-sum(rep(1,4)[ris[19:22]==2])
    t4[3]<-sum(rep(1,4)[ris[19:22]==3])
    t4[4]<-sum(rep(1,4)[ris[19:22]==4])

    t<-rbind(t1,t2,t3,t4)
    val<-NULL
    for(j in 1:24){
       ind<-perm[,j]
       val[j]<-
sum(t[1,ind[1]],t[2,ind[2]],t[3,ind[3]],t[4,ind[4]])
       }
    val
    te[i]<-(22-max(val))/22

    }
te
}
```

A5.2.r11

```
sr.k<-function(data,ga,gb,gc,gd,prob){
a1<-c(1,2,3,4,4,3,2,1,2,4,1,3,3,1,4,2,4,3,1,2,2,1,3,4)
a2<-c(2,3,4,1,3,2,1,4,4,1,3,2,1,4,2,3,3,1,2,4,1,3,4,2)
a3<-c(3,4,1,2,2,1,4,3,1,3,2,4,4,2,3,1,1,2,4,3,3,4,2,1)
a4<-c(4,1,2,3,1,4,3,2,3,2,4,1,2,3,1,4,2,4,3,1,4,2,1,3)
perm<-rbind(a1,a2,a3,a4)
te<-rep(0,length(prob))
data<-t(data)
ga<-t(ga)
gb<-t(gb)
gc<-t(gc)
gd<-t(gd)
mg<-apply(data,1,mean)
ma<-apply(ga,1,mean)
```

```
mb<-apply(gb,1,mean)
mc<-apply(gc,1,mean)
md<-apply(gd,1,mean)
ssa<-apply(((ga-ma)**2),1,sum)
ssb<-apply(((gb-mb)**2),1,sum)
ssc<-apply(((gc-mc)**2),1,sum)
ssd<-apply(((gd-md)**2),1,sum)
si<-sqrt((1/(ncol(data)-4))*(ssa+ssb+ssc+ssd))
so<-median(si)
m<-rep(0,4)
m[1]<-sqrt((1/ncol(ga))+(1/ncol(data)))
m[2]<-sqrt((1/ncol(gb))+(1/ncol(data)))
m[3]<-sqrt((1/ncol(gc))+(1/ncol(data)))
m[4]<-sqrt((1/ncol(gd))+(1/ncol(data)))
da<-(ma-mg)/(m[1]*(si+so))
db<-(mb-mg)/(m[2]*(si+so))
dc<-(mc-mg)/(m[3]*(si+so))
dd<-(md-mg)/(m[4]*(si+so))
da1<-abs(da)
db1<-abs(db)
dc1<-abs(dc)
dd1<-abs(dd)
dt<-cbind(da1,db1,dc1,dc1)
dt<-apply(dt,1,max)
delta<-rep(0,length(prob))


for(i in 1:length(delta)){
delta[i]<-quantile(dt,prob=prob[i])
p1<-abs(da)-delta[i]
p2<-abs(db)-delta[i]
p3<-abs(dc)-delta[i]
p4<-abs(dd)-delta[i]
dda<-parte.pos(p1)$y
ddb<-parte.pos(p2)$y
ddc<-parte.pos(p3)$y
ddd<-parte.pos(p4)$y

dda<-sign(da)*dda
ddb<-sign(db)*ddb
ddc<-sign(dc)*ddc
ddd<-sign(dd)*ddd

sel<-c(1:nrow(data))
sel<-sel[dda>0|ddb>0|ddc>0|ddd>0]
sel
dati<-t(data[sel,])
m1<-apply(dati[1:10,],2,mean)
m2<-apply(dati[11:15,],2,mean)
m3<-apply(dati[16:18,],2,mean)
m4<-apply(dati[19:22,],2,mean)
med<-rbind(m1,m2,m3,m4)

    ris<-kmeans(dati,med)$cluster
    t1<-rep(0,4)
    t2<-rep(0,4)
```

```
    t3<-rep(0,4)
    t4<-rep(0,4)
    t1[1]<-sum(rep(1,10)[ris[1:10]==1])
    t1[2]<-sum(rep(1,10)[ris[1:10]==2])
    t1[3]<-sum(rep(1,10)[ris[1:10]==3])
    t1[4]<-sum(rep(1,10)[ris[1:10]==4])

    t2[1]<-sum(rep(1,5)[ris[11:15]==1])
    t2[2]<-sum(rep(1,5)[ris[11:15]==2])
    t2[3]<-sum(rep(1,5)[ris[11:15]==3])
    t2[4]<-sum(rep(1,5)[ris[11:15]==4])

    t3[1]<-sum(rep(1,3)[ris[16:18]==1])
    t3[2]<-sum(rep(1,3)[ris[16:18]==2])
    t3[3]<-sum(rep(1,3)[ris[16:18]==3])
    t3[4]<-sum(rep(1,3)[ris[16:18]==4])

    t4[1]<-sum(rep(1,4)[ris[19:22]==1])
    t4[2]<-sum(rep(1,4)[ris[19:22]==2])
    t4[3]<-sum(rep(1,4)[ris[19:22]==3])
    t4[4]<-sum(rep(1,4)[ris[19:22]==4])


    t<-rbind(t1,t2,t3,t4)
    val<-NULL
    for(j in 1:24){
       ind<-perm[,j]
       val[j]<-
sum(t[1,ind[1]],t[2,ind[2]],t[3,ind[3]],t[4,ind[4]])
       }
    val
    te[i]<-(22-max(val))/22

    }
te
}
```

A5.2.r12

```
sr.pam<-function(data,ga,gb,gc,gd,delta){
a1<-c(1,2,3,4,4,3,2,1,2,4,1,3,3,1,4,2,4,3,1,2,2,1,3,4)
a2<-c(2,3,4,1,3,2,1,4,4,1,3,2,1,4,2,3,3,1,2,4,1,3,4,2)
a3<-c(3,4,1,2,2,1,4,3,1,3,2,4,4,2,3,1,1,2,4,3,3,4,2,1)
a4<-c(4,1,2,3,1,4,3,2,3,2,4,1,2,3,1,4,2,4,3,1,4,2,1,3)
perm<-rbind(a1,a2,a3,a4)
te<-rep(0,length(prob))
data<-t(data)
ga<-t(ga)
gb<-t(gb)
gc<-t(gc)
gd<-t(gd)
mg<-apply(data,1,mean)
ma<-apply(ga,1,mean)
mb<-apply(gb,1,mean)
mc<-apply(gc,1,mean)
md<-apply(gd,1,mean)
```

```
ssa<-apply(((ga-ma)**2),1,sum)
ssb<-apply(((gb-mb)**2),1,sum)
ssc<-apply(((gc-mc)**2),1,sum)
ssd<-apply(((gd-md)**2),1,sum)
si<-sqrt((1/(ncol(data)-4))*(ssa+ssb+ssc+ssd))
so<-median(si)
m<-rep(0,4)
m[1]<-sqrt((1/ncol(ga))+(1/ncol(data)))
m[2]<-sqrt((1/ncol(gb))+(1/ncol(data)))
m[3]<-sqrt((1/ncol(gc))+(1/ncol(data)))
m[4]<-sqrt((1/ncol(gd))+(1/ncol(data)))
da<-(ma-mg)/(m[1]*(si+so))
db<-(mb-mg)/(m[2]*(si+so))
dc<-(mc-mg)/(m[3]*(si+so))
dd<-(md-mg)/(m[4]*(si+so))
da1<-abs(da)
db1<-abs(db)
dc1<-abs(dc)
dd1<-abs(dd)
dt<-cbind(da1,db1,dc1,dc1)
dt<-apply(dt,1,max)
delta<-rep(0,length(prob))


for(i in 1:length(delta)){
delta[i]<-quantile(dt,prob=prob[i])
p1<-abs(da)-delta[i]
p2<-abs(db)-delta[i]
p3<-abs(dc)-delta[i]
p4<-abs(dd)-delta[i]
dda<-parte.pos(p1)$y
ddb<-parte.pos(p2)$y
ddc<-parte.pos(p3)$y
ddd<-parte.pos(p4)$y
dda<-sign(da)*dda
ddb<-sign(db)*ddb
ddc<-sign(dc)*ddc
ddd<-sign(dd)*ddd

sel<-c(1:nrow(data))
sel<-sel[dda>0|ddb>0|ddc>0|ddd>0]
sel
dati<-t(data[sel,])
    ris<-pam(dati,4,diss=F)$clustering
    t1<-rep(0,4)
    t2<-rep(0,4)
    t3<-rep(0,4)
    t4<-rep(0,4)
    t1[1]<-sum(rep(1,10)[ris[1:10]==1])
    t1[2]<-sum(rep(1,10)[ris[1:10]==2])
    t1[3]<-sum(rep(1,10)[ris[1:10]==3])
    t1[4]<-sum(rep(1,10)[ris[1:10]==4])

    t2[1]<-sum(rep(1,5)[ris[11:15]==1])
    t2[2]<-sum(rep(1,5)[ris[11:15]==2])
    t2[3]<-sum(rep(1,5)[ris[11:15]==3])
```

```
    t2[4]<-sum(rep(1,5)[ris[11:15]==4])

    t3[1]<-sum(rep(1,3)[ris[16:18]==1])
    t3[2]<-sum(rep(1,3)[ris[16:18]==2])
    t3[3]<-sum(rep(1,3)[ris[16:18]==3])
    t3[4]<-sum(rep(1,3)[ris[16:18]==4])

    t4[1]<-sum(rep(1,4)[ris[19:22]==1])
    t4[2]<-sum(rep(1,4)[ris[19:22]==2])
    t4[3]<-sum(rep(1,4)[ris[19:22]==3])
    t4[4]<-sum(rep(1,4)[ris[19:22]==4])


    t<-rbind(t1,t2,t3,t4)
    val<-NULL
    for(j in 1:24){
       ind<-perm[,j]
       val[j]<-
sum(t[1,ind[1]],t[2,ind[2]],t[3,ind[3]],t[4,ind[4]])
       }
    val
    te[i]<-(22-max(val))/22

    }
te
}
```

Capitolo 6
Considerazioni conclusive

Nel corso del presente elaborato sono stati analizzati dati riguardanti l'espressione genica, di ventidue pazienti affetti da quattro diverse forme di leucemia. L'analisi era finalizzata alla ricerca di tecniche in grado di riconoscere il tipo di leucemia di cui è affetto un paziente sulla base dei suoi livelli d'espressione genica nonché all'individuazione dei geni responsabili dell'insorgere della patologia e del suo differenziarsi in varie forme.

Dalle analisi condotte emerge che, dal punto di vista statistico, una corretta classificazione delle unità nelle quattro classi è realizzabile anche con le tecniche classiche (analisi discriminante lineare e logistica) e con l'utilizzo di un numero limitato di variabili (ossia geni) selezionate mediante criteri convenzionali come gli alberi di classificazione o il criterio di Akaike. Questi metodi consentono l'identificazione di una ventina di geni dotati di un alto potere discriminante. Sottoposti ad analisi biologiche, tali geni hanno però suscitato curiosità da parte dei ricercatori, in quanto per alcuni di questi non è ancora nota la funzione biologica. Comunque, nelle precedenti analisi da loro eseguite, alcuni di questi geni erano effettivamente già stati evidenziati in quanto dotati di caratteristiche discriminanti. Inoltre per quanto riguarda quelli di cui è nota la funzione biologica, si è constatato che tratta di geni coinvolti nella proliferazione oppure di oncogeni e oncosopressori.

Dal punto di vista biologico, si è posto l'accento sull'importanza dell'individuazione di (anche ampie) classi di geni responsabili dell'insorgere della patologia e/o del suo differenziarsi in varie forme. Sicuramente infatti all'interno di un complesso organismo biologico, non saranno solo una decina

geni che regolano, reagiscono o sono alterati da una patologia come la leucemia. La ricerca e la selezione di geni differenzialmente espressi è stato pertanto uno delle obiettivi di questo elaborato. Si sono utilizzate prevalentemente due tecniche proposte in letteratura una da Tibshisani e l'altra da Doudoit. I geni così selezionati sono poi stati analizzati dai ricercatori del CRIBI. Controllando la loro funzione biologica, sembra che essi siano coinvolti in classi funzionali (proliferazione cellulare, apoptosi, cell-interaction) tipiche delle patologie tumorali in genere, e quindi anche delle leucemie. Alcuni dei geni selezionati con il metodo di Tibshirani non erano emersi nelle precedenti ricerche e non erano stati fatti entrare nella selezione costituita dal *dataset* ridotto.

È stata inoltre sperimentata un'altra tecnica di selezione mediante le curve di Andrews e basata sul livello toccato dai profili generati dai geni per le diverse patologie. I geni selezionati sono stati usati cone variabili esplicative per l'analisi discriminante classica e per quella logistica, senza rivelare però un alto potere discriminante. Probabilmente tra i geni selezionati persiste ancora rumore. Infatti sembra che i geni evidenziati con questo metodo siano per il 25% geni differenzialmente espressi tra i gruppi ALL-B ALL-T ALL-Tr e AML. Nel corso del capitolo dedicato a questa tecnica si è comunque sottolineato che, per problemi di elaborazione, non è stato possibile condurre tale applicazione con un metodo più ottimale. Probabilmente, ciò avrebbe messo in luce qualcosa di più significativo o magari avrebbe eliminato i geni con limitato potere discriminante.

Fin dalle prime analisi è emerso inoltre il problema dell'elevata dimensionalità, questione che si impone sempre quando si tratta con questo genere di dati. Questo problema rende instabili le stime dei parametri degli errori *standard* ad essi associati nonché della matrice di varianze e covarianze, soprattutto quando il numero di osservazioni di cui si dispone è molto limitato come nel caso in esame. Come si è visto, in un primo momento la questione è stata "aggirata" mediante la selezione di variabili. In una fase successiva si è invece sperimentata una tecnica (regressione penalizzata) che fosse in grado di

trattare con *dataset* dalla dimensionalità così elevata senza una selezione preliminare di variabili. La regressione logistica penalizzata ha riportato risultati soddisfacenti dal punto di vista della classificazione tuttavia non ha portato alla luce risultati significativi dal punto di vista biologico in quanto è finalizzata unicamente alla classificazione e non alla selezione di geni.

Complessivamente si può affermare che alcune delle tecniche analizzate hanno portato a risultati abbastanza sensati anche al punto di vista biologico. I geni individuati sembrano significativi anche se sarebbero necessarie delle ricerche biologiche più dettagliate per avere evidenze più concrete. Questo comunque esula dagli interessi del presente elaborato che si propone come una ricerca tutt'altro che definitiva.

Molte questioni restano comunque aperte anche a causa degli elevati costi di queste nuove tecnologie che non consentono di disporre di un numero adeguato di osservazioni rapportato alla grande quantità di variabili molte delle quali non apportano informazione significativa ma aggiungono soltanto rumore. Questo tipo di problema fino a poco tempo fa non era molto sentito, ma in quest'ambito assume proporzioni consistenti. Infatti una condizione fondamentale richiesta per l'applicazione della maggior parte delle tecniche di analisi è quella di disporre di un numero considerevole di osservazioni rispetto alle variabili analizzate. Quando questa condizione non sussiste, molte tecniche non possono essere impiegate e per altre le stime dei parametri si fanno molto instabili.

Questa ed altre questioni danno ampio spazio alla ricerca e alla collaborazione tra diverse discipline. Molto probabilmente il continuo sviluppo tecnologico oltre che della biologia molecolare darà accesso, nei prossimi anni, a *dataset* più adeguati e a misure più attendibili, cosa che garantirà anche ricerca statistica delle basi più stabili su cui fondare il proprio lavoro.

Ringraziamenti

A conclusione del mio corso di studi desidero esprimere la mia gratitudine a quanti, in questi anni, mi hanno aiutato e sostenuto rendendo possibile la realizzazione di quello che all'inizio sembrava solo desiderio lontano.

In particolare ringrazio la mia relatrice, professoressa Monica Chiogna, per la disponibilità e la cortesia dimostratami in questo periodo di tesi e per aver messo a mio servizio la sua professionalità e le sue competenze.

Ringrazio di cuore la dottoressa e co-relatrice Chiara Romualdi e il dottor Nicola Sartori per la loro preziosa quanto gradita collaborazione.

Ringrazio i miei genitori che in questi anni non hanno mancato di incoraggiarmi sostenermi e consigliarmi, oltre che di assumersi gli oneri della mia istruzione.

Non sarò mai abbastanza riconoscente verso disponibilità di mio fratello Alessandro che, in particolar modo in questo periodo di tesi, ma anche in precedenza, ha messo a mia completa disposizione di tutte le sue attrezzature informatiche oltre che la sua indubbia competenza in materia, rendendo così la mia attività molto più comoda ed efficiente.

Ringrazio di cuore Christian che in tutti questi anni ha saputo sopportare e capire i miei cambiamenti d'umore, permettendomi di dedicarmi allo studio con la consapevolezza di avere accanto una persona in grado di comprendere rispettare i miei tempi.

Ringrazio di cuore Fabio De Min per avermi gentilmente e prontamente fornito le stampe a colori.

Ringrazio Elisabetta, oltre che per i ricordi che l'hanno legata a me in questi anni, per essersi resa disponibile a correggere e rivedere la parte biologica del mio elaborato.

Ringrazio Tanya per la comprensione, gli incoraggiamenti e l'interesse dimostrato in questi mesi nei confronti di ciò che stavo facendo.

Ricordo infine con tanto affetto e gratitudine tutti coloro che in questi anni mi hanno fatto dono della loro amicizia rendendo gradevole e felice la mia permanenza a Padova ed ogni ricordo ad essa legato. Con alcune di queste persone ho avuto modo di studiare, di apprendere e di approfondire le mie conoscenze, ricevendo gli stimoli per crescere e maturare anche intellettualmente: a tutti loro va il mio ringraziamento più sentito.

Martina Garlet

Riferimenti e bibliografia

[1] Gordon K. Smyth, Yee Hwa Yang and Terry Speed *Statistical Issues in cDNA Microarray Data Analysis* (2002)

[2] David M. Rocke and Blythe Durbin *A Model for Measurement Error for Gene Expression Arrays* (2001)

[3] B. Lausen *Statistical analysis of genetic distance data* (1999)

[4] Golub,T.R.,Slonim, D.K.Tamayo, P.,Huard et al (1999) *Molecular classification of cancer: class discoveri and class prediction by gene expression monitoring* Science 286:531-537

[5] Jean-Michel Clavarie *Computational methods for the identification of differential coordinated gene expression data* (1999)

[6] T. R. Golub, D. K. Slonim, Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh,J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander *Molecular Classification of Cancer: Class Discovery andClass Prediction by Gene Expression Monitoring* (Oct 1999)

[7] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Scho¨lkopf, C.J.C. Burges, A.J. Smola (Eds.), Advances in Kernel Methods—Support Vector Learning, MIT Press, Cambridge, MA, (1999), pp. 185–208.

[8] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T. Fury, M. Ares, D. Haussler, *Knowledge-based analysis of microarray gene expression data by using support vector machines*, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 262–267.

[9] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, Z. Yakhini, *Tissue classification with gene expression profiles*, J. Comput. Biol. 7 (2000) 559– 584.

[10] A.A Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A.Rosenwald, J.C. Boldrick, et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.

[11] Perou, CM, Serlie, T, Elsen, MB, van deRijn, M, Jeffrey, SS, Rees, CA, Pollack, JR, Ross, DT, Johnsen, H, Akslen, LA, et al. : *Molecular portraits of human breast tumours* Nature (2000), 406:747-752.

[12] van't Veer, LJ, Dai, H, vande Vijver, MJ, He, YD, Hart, AM, Mao, M, Peterse, HL, van derKooy, K, Marton, MJ, Witteveen, AT, et al. : *Gene*

expression profiling predicts clinical outcome of breast cancer Nature (2002), 415:530-536.

[13] Amir Ben-Dor , Laurakay Bruhn , Nir Friedman , Iftach Nachman , Michèl Schummer , Zohar Yakhini, *Tissue classification with gene expression profiles, Proceedings of the fourth annual international conference on Computational molecular biology*, p.54-64, (April 2000), Tokyo, Japan

[14] A. Keller, M. Schummer, L. Hood, W. Ruzzo, Bayesian classification of DNA *array* expression data, Technical Report, University of Washington, (August 2000)

[15] Andrew D. Keller, Michel Schummer Lee Hood Walter L. Ruzzo *Bayesian Classification of DNA Array Expression Data* Technical Report UW-CSE-2000-08-01 (August, 2000)

[16] Gen Hori, Masato Inoue, Shin-ichi Nishimura and Hiroyuki Nakahara *Blind gene classification based on ICA of microarray data*

[17] K. Zhang, H. Zhao, *Assessing reliability of gene clusters from gene expression data, Funct. Integr. Genomics* 1 (2000) 156–173.

[18] M.K. Kerr, G.A. Churchill, *Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiment*s, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 8961–8965.

[19] I. Steinwart. *On the influence of the kernel on the generalization ability of support vector machines*. Technical Report 01–01, (2001).

[20] Khan, J, Wei, JS, Ringner, M, Saal LH Ladanyi, M, Westermann, F, Berthold, F, Schwab, M, Antonescu, CR, Peterson, C, & Meltzer, PS: *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networ*ks Nature Medicine (2001), 7:673-679

[21] Dhanasekaran, SM, Barrette, TR, Chosh, D, Shah, R, Varambally, S, Kurachi, K, Pienta, KJ, Rubin, MA, & Chinnalyan, AM: *Delineation of prognostic biomarkers in prostate cancer* Nature 2001, 412:822-826.

[22] Allison, David B. , Gadbury, Gary L. , Heo, Moonseong , Fernández, José R. , Lee, Cheol-Koo , Prolla, Tomas A. , and Weindruch, Richard A mixture model approach for the analysis of micro*array* gene expression data", *Computational Statistics and Data Analysis* (2002), ``, 39 (1) , 1-20

[23] Lorenz Wenisch *Statistical methods for microarray data* (2001)

[24] Robert Tibshirani, Guenther Walther; David Botstein, Patrick Brown (2001) *Cluster validation by prediction strength*

[25] Therese Biedl, Brona Brejov, Erik D. Demaine, Angµele M. Hamel, and Tomas (2001) Vinar*Optimal Arrangement of Leaves in the Tree Representing Hierarchical Clustering of Gene Expression Data*

[26] Katherine S. Pollard Mark J. van der Laan (2001) *A Method to Identify Signi_cant Clusters in Gene Expression Data U.C. Berkeley Division of Biostatistics Working Paper Series*

[27] Sunduz Keles Mark J. van der Laany, Michael B. Eisenz Division (2001) *Identification of Regulatory Elements Using A Feature Selection Method*

[28] Jie Liang Serman Kachalo (2002) *Computational analysis of microarray gene expression profiles clustering, classification and beyond*

[29] Parmigiani, Giovanni , Garrett, Elizabeth S. , Anbazhagan, Ramaswamy , and Gabrielson, Edward *A statistical framework for expression-based molecular classification in cancer (2002) Journal of the Royal Statistical Society,* Series B, Methodological, 64 (4) , 717-736

[30] Fraley, Chris , and Raftery, Adrian E *Model-based clustering, discriminant analysis, and density estimation Journal of the American Statistical Association,* 97 (458) , 611-631 (2002)

[31] Darlene R. Goldtein, Debashis Ghosh and Erin M. Conlon (2002) *Statistical issues in the clustering of gene expression data Statistica sinica*

[32] M Kathleen Kerr, Cynthia A. Afshari, Lee Bennet, Pierre Bushel, Jeanelle Martinez, Nigel J. Walker and Gray A Churchill (2002) *Statistical analysis of gene expression microarray experiment with replication Statistica Sinica*

[33] Tibshirani, Robert , Hastie, Trevor , Narasimhan, Balasubramanian , Eisen, Michael , Sherlock, Gavin , Brown, Pat , and Botstein, David *Exploratory screening of genes and cluster from microarray experiment Statistica Sinica,* 12 (1) , 47-59 (2002)

[34] Jenny Bryan Katherine s. Pollard Mark J. Van der Laan (2002) *Paired and unpaired comparison and clustering withn gene expression data Statistica Sinica*

[35] Laura Lazzeroni and Art Owen (2002) *Plaid model for gene expression data Statistica sinica*

[36] Gengxin Chen, Saied A. Jaradat, Nila Banerjee (2002) *Evalutation dna comparison of clustering algorithms in analyzing es cell gene expression data Statistica sinica*

[37] Doudoit S, Fridlyand J and Speed T.P. (2002) *comparison of discrimination method for identifying differentially expressed genes in replicated cDNA microarray experiment* Statistica sinica 12: 111-139

[38] Bradley Reflon and Tibshirani *Empirical bayes mrthods and false discovery rates for microarray* (2002)

[39] Francesca Ruffino, Giorgio Valentini e Marco Muselli *Metodi di Bagging e di selezione delle variabili per l'analisi dei dati di DNA microarray* (2002)

[40] Chris H.Q. Ding *Analysis of gene expression profiles: class discovery dna leaf ordering* (2002)

[41] Katherine S. Pollard Mark J. van der Laan *A Method to Identify Signi_cant Clusters in Gene Expression Data U.C. Berkeley Division of Biostatistics Working Paper Series* (2002)

[42] Chiara Romualdi, Stefano Campanaro, Davide Campagna; barbara Celegato, Nicola Cannata, Stafano Toppo, Giorgio Valle and Gerolamo Lanfranchi *Pattern recignition ingene expression profilino usung DNA array: a comparative study of different statistical methods applied to cancer classification Human molecolar genetics* (2003*) Vol 12. No 8*

[43] Margaret S. Pepe, ary M. Longtony, Garnet L. Anderson and Michel Schummer *Selecting Differentially Expressed Genes from Microarray Experiments UW Biostatistics Working Paper Series* (March 2003)

[44] Erich Huang, Skye H Cheng, Holly Dressman, Jennifer Pittman, Mei Hua Tsou, Cheng Fang Horng, Andrea Bild, Edwin S Iversen, Ming Liao, Chii MingChen, Mike West, Joseph R Nevins, Andrew T Huang *Gene expression predictors of breast cancer outcomes* (March 2003)

[45] Michel C O'Neill and Li Song *Neutral network of limphoma microarray data: prognosis and diagnosis near-pertfect* (April 2003)

[46] Susmita Datta and Somnath Datta Comparisons and validation of statistical *cluster*ing tecniques for micro*array* gene expression data *Bioinformatics* (2003)

[47] Paola Sebastiani, Emanuela Gussoni, Isaac S. Kohane and Marco F.Ramoni *Statistical Challenger in functional genomics* (2003)

[48] M. K. Kerr *Linear Models for Microarray Data Analysis: Hidden Similarities and Differences* (2003)

[49] Richard C. Dubes and Anil K. Jain, (1988), *Algorithms for Clustering Data,* Prentice Hall.

[50] L. Kaufman and P. J. Rousseeuw, (1990), *Finding Groups in Data: an Introduction to Cluster Analysis,* John Wiley and Sons.

[51] O. Troyaskaya, M.Cantor, G.Sherlock, P. Brown, T. Hastie, R. Tibshirani, D Botstein and R.B. Altman (2001) *Missing value estimation methods for DNA microarray* Bioinformatics

[52] R. Tibshirani, T, Hastie B. Narasimhan and G. Chu *Diagnosis of multiple cancer types by shrunken centroids of gene expression* (2001)

[53] Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classifcation and Regression Trees,* Chapman and Hall, New York, NY.

[54] Hyunjoong Kim and Wei-Yin Loh Classifcation Trees With Unbiased Multiway *Splits J. Amer. Statist. Assoc.*, (2001), 96, 598(604)

[55] Nagendra Kumar and Andreas G. Andreou *On Generalizations of Linear Discriminant Analysis* (1960) Report JHU/EC

[56] W. J. Krzanowski; P. Jonathan; W.V. McCarthy; M.R. Thomas Discriminant Analysis with Singular Covariance Matrices: Methods and Application to Spectroscopic Data Applied Statisctcs Vol. 44, No. 1 (1995), 101-115