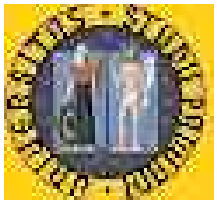


Università degli Studi di Padova
Facoltà di Scienze Statistiche

CORSO DI LAUREA TRIENNALE IN STATISTICA E TECNOLOGIE INFORMATICHE

Tesi di laurea triennale

ANALISI DELLE ESPRESSIONI GENICHE PER LA CLASSIFICAZIONE DI DIVERSI TIPI DI LEUCEMIE

RELATORE: CHIAR.MA PROF. MONICA CHIOGNA

CANDIDATA: LAURA GAVAGNIN

ANNO ACCADEMICO 2003/2004

Indice

Introduzione

Ormai da tempo, lo studio del DNA umano da parte di studiosi e scienziati sta prendendo piede in quanto, dopo una lunga analisi delle espressioni genetiche, si è ipotizzata la teoria secondo la quale gran parte delle malattie derivano da alterazioni del codice genetico.

La tecnologia del *DNA microarray* può essere un utile strumento per identificare mutazioni presenti nei geni e per comprendere, attraverso l'analisi simultanea di migliaia di geni, la patogenesi sia delle malattie genetiche vere e proprie (quelle cioè che si trasmettono in famiglia), sia di quelle multifattoriali come diabete, osteoporosi, arteriosclerosi. Questa strumentazione offre quindi un ottimo mezzo per cinque principali obbiettivi biologici:

1. l'identificazione di geni con livelli di espressione diversa sotto diverse condizioni sperimentali o tra soggetti che presentano varie forme della stessa patologia;
2. l'identificazione di gruppi di geni che con buona probabilità sono correlati tra loro;
3. la caratterizzazione gnomica della cellula malata attraverso la classificazione di campi biologici (soggetti sani vs soggetti affetti da una determinata patologia);
4. l'identificazione di geni il cui valore di espressione è biologicamente utile per determinare un particolare gruppo o fenotipo (tali geni sono detti marcatori);
5. l'identificazione di nuove classi di una specifica patologia (come nel caso delle patologie oncologiche: esisto molte classi di tumori diversi).

L'identificazione di mutazioni è fondamentale per la prevenzione delle malattie genetiche, per la diagnostica precoce dei tumori, nonché in microbiologia per la identificazione di ceppi batterici o virali. Un altro settore di applicazione è quello dell'analisi funzionale simultanea di decine di migliaia di geni e, in un futuro prossimo, di tutti i geni che costituiscono il nostro patrimonio genetico. Inoltre i risultati che ci si aspetta di ottenere con questa nuova tecnologia saranno fondamentali per sviluppare nuovi farmaci, e per meglio utilizzare quelli attualmente disponibili dando al medico la possibilità di adattare la terapia sulla base delle caratteristiche genetiche di ognuno di noi.

Gli studi sul cancro costituisco 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. E' 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 la 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 manipolare simultaneamente migliaia di geni, motivando lo sviluppo della classificazione di tumore con l'utilizzo dei dati d'espressione genica.

Nel presente elaborato ci si propone lo studio delle leucemie soprattutto nell'ottica della classificazione. Lo scopo sarà pertanto quello di distinguere tra tre gruppi di leucemie sulla base dell'espressione genica di migliaia di geni.

Lo scopo dell'analisi è quello di individuare una regola di classificazione che permetta di allocare una nuova osservazione (individuo) ad una delle tre popolazioni (tipi di leucemia), in base ai valori assunti dalle variabili esplicative (espressioni geniche). In termini statistici si tratta di costruire un modello che preveda al meglio la possibile leucemia partendo dalle p variabili esplicative; non interessa dunque un buon adattamento del modello ai dati, bensì la capacità di previsione dello stesso.

Occorre individuare, quindi, una regola che sbagli il meno possibile in termini previsivi, costruendo un modello che non dipenda troppo dal campione di dati che l'ha generato.

Nello specifico si utilizzerà regressione logistica penalizzata che bene si adatta a questa tipologia di dati. A differenza della regressione logistica classica, nella regressione logistica penalizzata è presente un parametro – detto appunto penalizzazione – che controlla la complessità del modello. Il parametro di penalizzazione ha un'importanza fondamentale ma nella realtà, nella letteratura statistica non è ancora accuratamente sviluppato.

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*. rif[1] hanno stilato una rassegna dei problemi statistici.

Il primo problema che pone il *dataset* è la 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 del modello tanto per l'alta dimensionalità quanto per la limitata numerosità campionaria. In secondo luogo la maggior parte dei geni nel *dataset* sono irrilevanti ai

fini della classificazione e costituiscono un *rumore* che interferisce con il potere discriminante degli altri geni. Questo accresce non solo i tempi di calcolo, ma anche la difficoltà di classificazione. E' evidente che i metodi di classificazione esistenti non sono concepiti per essere applicati a questo tipo di dati. Alcuni ricercatori propongono il raggruppamento di geni in classi omogenee come operazione preliminare alla classificazione dei soggetti, in quanto tale operazione ha la capacità di ridurre la dimensionalità, i tempi di calcolo ed eliminare i geni irrilevanti che comportano minor accuratezza nella classificazione. Una terza questione riguarda la natura stessa dei dati che sono caratterizzati dalla massiccia presenza di rumore di tipo *biologico* o *tecnico*.

I problemi sopra menzionati riguardano l'ambito statistico, ma esistono diverse questioni derivanti dal contesto biologico e dall'importanza dei risultati in campo medico. Una questione riguarda la corrispondenza tra rilevanza biologica e statistica di uno stesso gene come classificatore: la rilevanza biologica è un criterio da tenere in forte considerazione in quanto ogni informazione rilevata durante l'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 cancerogeni, per la scoperta di interazione tra i geni o per altri studi biologici come l'individuazione di geni marcatori. Infine esiste una questione chiamata *contaminazione del campione*: usualmente tessuti normali e cancerogeni sono composti da cellule differenti, il tessuto tumorale è ricco di cellule epitali mentre il tessuto normale è formato da una grossa porzione di cellule muscolari. Questo può condurre ad una selezione di geni che hanno diversi valori di espressione nei due tessuti, ma tale differenza è imputabile ad una diversa composizione dei tessuti stessi con il risultato di una buona regola di classificazione ma senza fornire risultati biologici rilevanti.

La letteratura non è molto datata e tende a svilupparsi a pari passo con le scoperte in ambito biologico.

Nel 2001 Rocke e Durbin rif [2] introducono un modello di misura per gli errori dei dati da *microarray* come funzione del livello di espressione dei geni.

Nel 1999 Lausen rif [3] si concentra sulle misure di distanza allineando sequenze di dati secondo diversi criteri, propone poi un grafico (*dot-matrix plot*) come possibile test sulla bontà dell'allineamento. 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. Jean Clavarie rif [5] rivede invece l'approccio teorico e computazionele utilizzato fino ad allora per identificare i geni differenzialmente espressi, per selezionare 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 Gloub *et al* rif [6] applicano due procedure di classificazione (*class discovery* e *class prediction*) per distinguere diversi tipi di cancro per leucemie acute. Platt rif [7]

mette appunto la *"sequential minimal optimisation"* che permette l'implementazione delle *support vector machines* (SVM) per affrontare problemi di classificazione che coinvolgono grandi *dataset*.

L'anno successivo Brown *et al* rif [8] testano diverse SVM usando varie misure di sorveglianza su dati da microarray trovando le SVM garantiscono prestazioni migliori rispetto ad altre tecniche nel riconoscere geni coinvolti nelle comuni funzioni biologiche. Ben Dor (2000) *et al* rif [9] descrivono un'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 regole di raggruppamento gerarchico per studiare l'espressione genetica nelle tre prevalenti forme di tumore linfoide che colpisce gli adulti. 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 dei *cluster* in cui raggruppano geni simili. Veer *et al* (2000) rif [12] studiano un *dataset* di 78 pazienti con il 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 Bem-Dor *et a*l rif [13] verificano che mentre le SVM hanno maggiore accuratezza sui dati da leucemie e i metodi basati sul *clustering* funzionano meglio su dati di tumori al colon, il metodo *nearest neighbor* dà buoni risultati in entrambi i casi. Keller *et a*l (2000) rif [14] comparano il metodo bayesiano semplice con il metodo *weighted voting* e nell'agosto dello stesso anno presentano il primo metodo per la classificazione di tipi di tessuto con dati da microarray usando una tecnica basata sulla massima verosimiglianza per selezionare i geni più utili alla classificazione. Applicando questa tecnica ad un *dataset* con due tipi di tessuti riscontrano un'eccezionale accuratezza e fanno notare che è facilmente estendibile ad una classificazione con più di due classi fornendo ottimi risultati se applicati a *dataset* con tre tipi di tessuto. Gen Hori *et al* (2000) rif [15] 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 [16] e Kerr *et al* rif [17] usano il metodo *bootstrap* per valutare la qualità dell'analisi di raggruppamento i primi assumendo che i livelli di espressione hanno distribuzione normale, i secondi usando il modello ANOVA per generare campioni *bootsrapped*

Nel maggio del 2001 David B., Allison *et al* rif [18] sviluppano una sequenza di procedure che comprendono modelli misti e inferenza *bootstrap* per affrontare problemi (come *large p and small*

n) che sorgono nel trattamento dei dati che coinvolgono l'espressione di migliaia di geni. Nel luglio dello stesso anno Lorenz Wernisch rif [19] propone una rassegna dei principali metodi di trattamento dei dati da microarray. Tibshirani *et al* (2001) rif [20] propongono una quantità per la stima del numero di *cluster* in un *dataset*: tale 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 variabile risposta.

Nel 2002 Chris Fraley e Adrian E. Raftlery rif [21] 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 (*oulyers*), 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. Sempre nello stesso anno Gengxin Che *et al* rif [22] 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 *silouette*, sui geni in eccedenza in un dato gruppo ecc. I risultati dimostrano che il dataste pone effettivamente dei problemi per l'analisi *cluster*, gli autori valutano vantaggi e svantaggi dei vari algoritmi. Lo studio fornisce quindi una linea generale su come scegliere tra diversi algoritmi e può aiutare ad estrarre dal *dataset* le informazioni biologiche più significative.

Nel febbraio del 2003 Romualdi, Campanaro *et al* rif [23] comparano diverse tecniche di *supervised clustering* 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 *dataset* di espressioni geniche di pazienti malati di cancro, misurando il corrispondente tasso di errata classificazione. Nel marzo dello stesso anno Erich Hungan *et al* rif [24] 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 accuratezza del 90%. Nell'aprile del 2003 Michael O'Neil e Li Song rif [25] 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 miglior strumento sia per l'individuazione di gruppi di geni sia per evidenziare i geni più importanti che producono una corretta classificazione.

Capitolo 1

I Microarray

1.1 Breve rassegna delle scoperte riguardanti la molecola di DNA

Ogni essere vivente possiede un programma genetico, cioè un insieme di istruzioni che specificano le sue caratteristiche e dirigono le sue attività metaboliche. Questo insieme di istruzioni costituisce l'informazione biologica, cioè è ereditaria ed è trasferita da una generazione all'altra attraverso la riproduzione. Le caratteristiche trasmesse sono dette caratteri ereditari.

L'informazione biologica è organizzata in unità fondamentali, dette geni, ciascuna delle quali interviene nella determinazione di un carattere ed è ereditata dai genitori.

Già con le prime ipotesi riguardanti l'evoluzione, si era cercato di comprendere come i caratteri peculiari di un organismo venissero trasmessi e come le specie evolvessero. Alcuni, come il noto scienziato Lamarck (1787), avevano ipotizzato che i caratteri acquisiti durante la vita fossero trasmissibili di padre in figlio: è il caso del famoso esempio della giraffa e del suo collo. Molte furono però le critiche rivolte a questa teoria, dovute al fatto che molti caratteri acquisiti durante la vita non sono ereditari.

Il primo a dedicarsi con metodo scientifico allo studio dell'ereditarietà dei caratteri fu un abate austiaco Gregor Mendel, Bateson W. (1909). A quei tempi Mendel non aveva nessuna conoscenza della struttura intrinseca del DNA, tuttavia aveva intuito alcuni caratteri che ricomparivano regolarmente nelle popolazioni. Le regole fondamentali che mettevano in connessione questi eventi non erano ancora chiare. I primi esperimenti che Mendel condusse furono sul piante di piselli odorosi caratterizzate dalla capacità di effettuare l'autofecondazione e caratterizzate da cicli vitali non troppo lunghi.

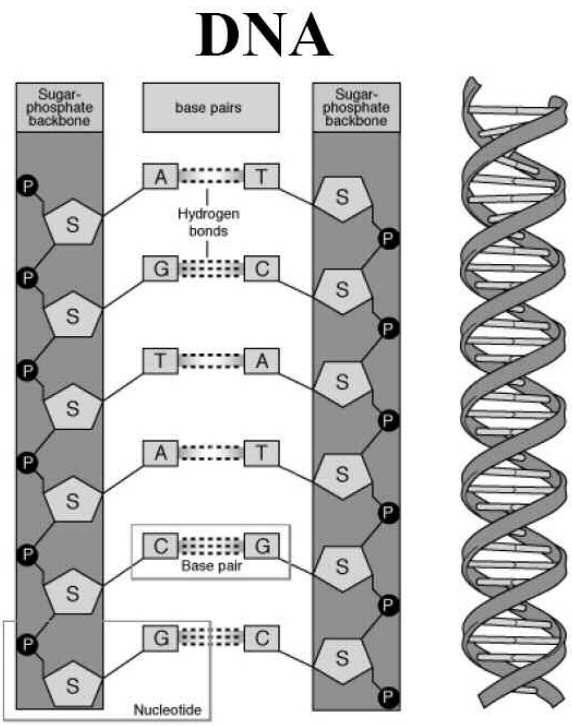
Le ricerche di Mendel non furono prese immediatamente in considerazione, ma le basi della genetica erano comunque scoperte e senza avere idea di come fosse strutturato il DNA.

La prova decisiva che il depositario dell'informazione è il DNA fu fornita nel 1952 da A.D. Harshey e M.Chase, i quali dimostrarono che i batteriofagi, per introdurre nel batterio ospite il loro materiale ereditario, iniettano una molecola di DNA.

L’importante scoperta fatta sul DNA suscitò la curiosità degli scienziati sulla struttura di tale molecola. Agli inizi degli anni ’50, un giovane scienziato americano, James Watson, si recò a Cambridge, in Inghilterra, con una borsa di studio per lavorare sui problemi di struttura molecolare e, al Cavendish Laboratory, incontrò il fisico Francis Crick. Entrambi si interessavano di DNA e ben presto cominciarono a lavorare insieme per cercare di capire come fosse strutturata tale molecola. Essi non eseguirono veri e propri esperimenti, ma intrapresero, piuttosto, un esame razionale dei dati allora noti sul DNA, cercando di organizzarli in modo logico. Le informazioni che essi avevano su tale molecola riguardavano le sue grosse dimensioni e la sua struttura lunga e filiforme formata da nucleotidi. Inoltre nel 1950 Linus Pauling aveva dimostrato che le proteine sono spesso disposte in maniera elicoide e vengono mantenuti in questa disposizione da legami idrogeno-idrogeno che si formano sulle spire adiacenti all’elica. Questa dimostrazione risultò utile ai fini della ricerca in quanto la molecola di DNA si comporta in modo simile alla molecola delle proteine.

Studi intrapresi da Maurice Wilkins e Rosalind Frankling ai raggi X dimostrarono la forma a grande elica del DNA. Infine Chargaff verificò l’esattezza di due proporzioni che dimostravano l’impossibilità di legare chimicamente due basi purine (a due anelli) o due basi pirimidine (ad una unico anello), e quindi all’assunzione che la timina si può legare solamente alla adenina, e la citosina solamente alla guanina.

L’insieme di tutte queste scoperte portò la formulazione della struttura definitiva della molecola di DNA: doppia elica lunga e spiralizzata in cui le due spirali sono formate da molecole alternate di zucchero e di fosfato e vengono tenute insieme da una coppia di basi azotate (ogni base è legata in modo covalente alla subunità glucidica posta nel tratto montante adiacente ad essa).



Sulla molecola di DNA sono state formulate numerose ipotesi e nel corso degli anni si è scoperto quasi con completezza come essa trasferisce l'informazione biologica da un individuo all'altro.

Numerosi scienziati si sono occupati e si occupano di individuare le particolarità del DNA tramite lo studio descrittivo dei geni e la loro classificazione.

L'applicazione più interessante è nello studio di malattie: molti studiosi concordano oramai da tempo sulla teoria secondo la quale alcune patologie derivino da piccole alterazioni del codice genetico. Ciò che distingue un individuo sano da un malato, sono delle differenze nell'espressione dei geni, ossia nel modo con cui essi sono utilizzati e nelle proteine a cui danno origine.

La disciplina metodologica che si occupa di questi problemi è la biostatistica la quale assiste il ricercatore biologo nel disegno e nella valutazione probabilistica di variazione di espressione genetica. Il problema è quello di caratterizzare le anomalie genetiche della cellula malata, ossia ciò che la differenzia da quella sana, in modo tale che una volta noto il profilo genetico di un paziente, risulti possibile identificarlo come sano o affetto da malattia.

1.2 Alcune nozioni di biologia

Per comprender 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 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 ed 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 costituitiva 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 ponti infinitesimali detti *ponti idrogeno*. I due filamenti sono costituiti 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 costituisco un *nucleotide*.

Esistono quattro tipi di basi: *adenina*, timida, *citosina*, *guanina* ed hanno la caratteristica di accoppiarsi sempre nello stesso modo, adenina con timida 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 le triplette, si ha la formazione di 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. E' 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 una sequenza degli amminoacidi, ossia delle subunità necessarie per la sintesi proteica: una serie di tre nucleotidi (detta *codone*) codifica per un amminoacido.

Il processo secondo la quale il DNA viene tradotto in proteine consta in 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, la sola differenza è che al posto della timida si trova un'altra base azotata: 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 subunità formate da RNA ribosomiale e proteine (o rRNA). A questo punto interviene l'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 allea 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 mutazione dei 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 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 Le nuove tecnologie: i microarray

Le nuove tecnologie di studio del DNA, *microarray*, descritte per la prima volta nel 1995, stanno rapidamente trovando applicazione in molti ambiti di ricerca, che vanno dalla fisiologia cellulare, all'oncologia, alla farmacogenomica. Numerose sono anche le applicazioni di questa tecnologia in ambito microbico-virologico, come la genotipizzazione e lo studio della biologia dei microrganismi e delle interazioni ospite-patogeno. Il sistema di indagine basato sui *microarray* permette di misurare contemporaneamente molte sequenze diverse, e quindi di analizzare l'intero patrimonio genetico di diversi organismi.

Attualmente sono state sviluppate due principali piattaforme tecnologiche per la produzione dei *microarray*:

- *Microarray di cloni di DNA micropipettati:* diversi singoli geni vengono depositati in anticipo sui vetrini opportunamente trattati con agenti chimici che favoriscono il legame del DNA utilizzando apparecchiature automatizzate;
- *Microarray di oligonucleoti disintetizzati in situ:* utilizza chip di silicio su cui sono direttamente sintetizzati oligonucleotidi rappresentativi della sequenza bersaglio.

I vantaggi dei primi sono l'alta automazione delle procedure sperimentali, una elevata riproducibilità dovuta al costo relativamente basso e il fatto che non sia necessario conoscere la sequenza del DNA da stampare; mentre i vantaggi dei microarray di oligonucleotidi sono l'alta densità e l'opportunità di disegnare la sequenza bersaglio dall'utente e quindi adattato alle diverse situazioni sperimentali.

Il *microarray* consente di verificare quanti e quali geni sono attivi in un tipo cellulare o in un tessuto, qual è il loro livello di espressione e quali variazioni accadono in condizioni patologiche. In tal modo e' possibile identificare i geni con potenziale attività oncogena che sono attivi nelle cellule tumorali di un paziente rispetto ad un altro, o rispetto al tessuto normale. Allo stesso modo si possono valutare quali geni differenziano il tumore primario dalla relativa metastasi.

Tutto ciò, oltre a costituire un ulteriore approccio sperimentale per l'identificazione di geni collegati al fenomeno della trasformazione e progressione neoplastica, ha permesso di classificare i tumori in base ai loro profili di espressione e di preparare lo sviluppo di una tassonomia molecolare dei tumori che consenta di complementare, aggiungendo nuove e più rilevanti informazioni, quella tradizionale di tipo isto-morfologico.

Le innovazioni che hanno reso possibile la tecnologia dei *microarray* sono l'uso di supporti solidi non porosi come vetro, molto versatile ai fini della miniaturizzazione e dell'individuazione dei marcatori fluorescente, e la sintesi ad alta densità spaziale di oligonucleotidi su vetrini sottilissimi con tecniche che utilizzano maschere fotolitografiche, impiegate nell'industria dei

semiconduttori. Tra le numerose applicazioni della tecnologia dei microarray, le principali sono l'analisi su larga scala dell'espressione genetica e la ricerca di variazioni della sequenza del DNA.

La tecnologia basata sui *microarray*, rappresenta un mezzo di indagine straordinariamente innovativo, in quanto permette di analizzare con un singolo esperimento l'intero patrimonio genetico di un organismo.

1.3.1 Come funziona l'esperimento

La realizzazione del *microarray* consta in due fasi: la preparazione del *microchip* e quella del *target*.

Ad un vetrino (*microchip*) si fissano delle sonde (*probe*) costituiti da segmenti di cDNA sintetico che riproducono i geni che in qualche modo sono notoriamente correlati con la patologia oggetto di studio. A questo scopo esistono speciali robot in grado di dispensare goccioline dell'ordine di nanolitri attraverso tubi con 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 retrotrascitto dando luogo al *cDNA* che, come ricordato precedentemente, presenta una molecola più stabile *dell'mRNA*. Durante questa fase nella catena di *cDNA* di ciascun gene vengono introdotte particolari molecole dette recettori 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 ricettori: Cy3 (verde) per 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'ibridazione i segmenti di *cDNA* target riconoscono le sonde complementari e si legano ad esse.

Una volta completata l'ibridazione il *microchip* viene levato e successivamente eccitato con un laser affinché i marcatori fluorescenti emettano un segnale luminoso. Una specie di *scanner* legge l'*array* illuminando ciascuno *spot* (ossia ciascun puntino che rappresenta 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 di *cDNA* contrassegnato con Cy3, e quindi *mRNA* prodotto da cellule sane; mentre quella degli spot rossi misura la quantità relativa di *cDNA* contrassegnato con Cy5, e quindi di *mRNA* prodotto da 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 o 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 cellula, ed infine il verde ad un gene attivo nella cellula sana e inattivo in quella malata.

E' necessario che queste misure vengano aggiustate per considerare un disturbo di fondo causato ad esempio dall'alta concentrazione di sale e detergente durante l'ibridazione o la contaminazione del target o da altri problemi che si possono presentare nell'esecuzione dell'esperimento.

L'ibridazione 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 lìimportanza della qualità dell'immagine ai fini di un'accurata interpretazione dei dati. I passi principali delle immagini prodotte da *cDNA microarray* sono:

1. grigliatura (*gridding*)
2. estrazione di intensità
3. segmentazione

La grigliatura ritrova nell'immagine la posizione degli spot che corrispondono alle sonde. Essendo nota la posizione degli spot nel microarray, questa operazione non risulta particolarmente complessa, sebbene si renda necessaria la stima di alcuni parametri per tener conto ad esempio di *shift* (o rotazioni) del *microarray* nell'immagine o di piccole traslazioni degli spot.

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

La segmentazione consiste infine nel separare il segnale emesso dai marcatori fluorescenti (*foreground*) rispetto al disturbo di fondo (*background*), in modo da isolare le quantità di interesse.

Può succedere che questa correzione abbia l'effetto indesiderato di introdurre valori negativi(ciò accade quando l'intensità del background è più forte rispetto a quella di foreground). In tal caso questi spot vengono trascurati oppure il loro segnale è sostituito con un valore arbitrariamente piccolo e positivo.

1.3.2 Tipi di distorsioni

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 e i più noti sono:

1. *dye-effect* (o effetto colore);
2. *print-tip* (o deposito irregolare);
3. *array-effect*(o effetto intensità).

Ad esempio, un diffuso problema nell'interpretazione dei dati derivanti da *microarray*, 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*, è dovuta alla diversa quantità di materiale genetico (probe) depositata sul vetrino a causa delle microscopiche differenze della conformazione delle puntine del rabor che stampa l'*array*.

Infine, il 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 (i) media generale di *array*, la (ii) differenza tra le medie degli spot stampati da ciascun *print-tip* e la media generale, ed infine la (iii) 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 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 di espressione costante e possono essere usati come indicatori dell'intensità relativa ai due colori. In alti termini, quasi tutti i geni dell'array possono essere utilizzati 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

trarre un fattore di correzione. Si preferisce quindi, quando il numero di geni differenzialmente espressi è limitato rispetto al numero totale dei geni indagati, usare tutti gli spot dell'array nel processo di normalizzazione dei dati. Il secondo approccio si basa sull'assunto che da proporzione di geni differenzialmente espressi sia un'altra 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 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.

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 viene raggiunto rimuovendo quei geni le cui misure non sono sufficientemente accurate, il secondo con l'imitazione dei geni che prevedono un livello di espressione molto piccolo o negativo (prima o dopo la normalizzazione).

In pratica, tutti gli spot la cui differenza tra l'intensità di foreground e quella di background non supera un valore soglia di 1.4 fold (una misura dell'intensità luminosa) vengono eliminati o sostituiti con un valore piccolo arbitrario. Questa procedura è giustificata dall'evidenza empirica che livelli di espressione più piccoli di 1.4 fold sono solitamente frutto di errori di misura.

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.

Capitolo 2

I dati

2.1 Il dataset

Il *dataset* preso in considerazione è stato fornito dal sito internet *http//dmc.org.sg/GEDataset/Datasets.html*. Contiene i profili genetici di 72 soggetti distinti in tre gruppi a seconda del tipo di leucemia da cui risultano affetti. Si hanno dunque 24 soggetti sono affetti da leucemia linfoblastica acuta (ALL), 20 sono affetti da leucemia mieloide lieve (MLL) e 28 da leucemia mieloide acuta (AML).

Per ciascun soggetto sono state raccolte le espressioni di 12582 geni. Viene infine fornita una divisione del dataset in una parte dedicata alla stima dei parametri per la classificazione (*training set*), e l'altra per la verifica della bontà della classificazione (*test set*).

Per lo scopo del nostro studio ci occuperemo di questi dati nei termini di classificazione dei soggetti nelle 3 leucemie, quindi le unità statistiche sono rappresentate dai vari soggetti e le variabili esplicative sono le espressioni dei geni.

Da una prima osservazione si può notare che non ci sono valori mancanti, quindi si possono mantenere tutte le unità statistiche. Per risolvere il problema dell'elevata numerosità esiste la possibilità di utilizzare procedure di classificazione che prevedono la selezione manuale delle variabili quali, per esempio, gli alberi di classificazione. Questi metodi però hanno la peculiarità di isolare pochi geni con elevato potere discriminante. Tale parsimonia non è molto gradita in biologia: i ricercatori infatti sono interessati ad individuare gruppi, anche piuttosto numerosi, di geni responsabili della diffusione della patologia e del suo differenziarsi in varie forme. Sono stati sperimentati a tale proposito alcuni metodi di selezione automatica e di controllo della dimensionalità attraverso l'utilizzo di parametri di penalizzazione.

2.2 Il controllo della dimensionalità

Il metodo proposto da Tibshirani e coautori rif [38] consente di evidenziare gruppi di espressioni geniche che contribuiscono in maniera più evidente alla classificazione delle patologie.

Tale criterio è una modificazione e reinterpretazione del metodo del centroide più vicino (*nearest-centroid*) utilizzato nell'analisi di raggruppamento. La sostanziale differenza è l'utilizzo dei centroidi di gruppo che si discostano di più dalla media generale, ovvero quelli per cui la differenza dal centroide generale supera una certa soglia regolata da un parametro di penalizzazione (Δ) fissato in modo da minimizzare l'errore derivante dalla validazione incrociata.

Sia x_{ij} l'espressione dell' i-esimo gene, $i = 1,\ldots,p$, nel j-esimo individuo, $j=1,\ldots,n$. Abbiamo inoltre $1,\ldots,K$ classi; indicheremo quindi C_k la classe k composta di n_k elementi. L' i-esimo componente del centroide per la k-esima classe è dato da $\bar{x}_{ik} = \sum_{j \in C_k} \frac{x_{ij}}{n_k}$, ovvero dalla media delle espressioni dell'i-esimo gene nella k-esima classe (media di gruppo); l'i-esimo componente del centroide generale è dato invece da $\bar{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n}$ ovvero dalla media generale.

Sia infine

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k (s_i + s_0)} \qquad (1)$$

dove s_i è la radice quadrata della somma delle varianze di classe per l'i-esimo gene; cioè $s_i = \left(\frac{1}{n-K} \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ij})^2 \right)^{1/2}$, e $m_k = \sqrt{\frac{1}{n_k} + \frac{1}{n}}$. La quantità d_{ik} rappresenta la statistica t per l'i-esimo gene per un sistema di ipotesi che confronta la k-esima classe con il centroide generale. Le quantità m_k e s_k sono definite in modo che $m_k * s_i$ sia pari alla stima dell' errore standard del

numeratore di d_{ik}. Nel denominatore, il valore s_0 è una costante positiva (con lo stesso valore per ogni gene) pari alla mediana di s_i che viene inclusa per evitare che geni con basso livello di espressione, generino valori elevati di d_{ik}.

Possiamo riscrivere l'equazione (1) come:

$$\bar{x}_{ik} = \bar{x}_i + m_k (s_i + s_0) d_{ik} . \qquad (2)$$

Il metodo proposto da Tibshirani consiste nel confrontare i valori d_{ik} con le soglia Δ producendo in questo modo i coefficienti:

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

dove $(\bullet)_+$ indica la funzione *parte positiva*:

$$t_+ = \begin{cases} t & se\ t > 0 \\ 0 & altrimenti. \end{cases},$$

Questi coefficienti servono infine per la selezione degli "***shrunken centroids***" definiti come:

$$\bar{x}'_{ik} = \bar{x}_i + m_k (s_i + s_0) d'_{ik} . \qquad (3)$$

Risulta evidente che, all'aumentare di Δ, diminuisce il numero di geni coinvolti nella regola di allocazione.

Per la scelta del parametro Δ ci si serve della validazione incrociata (*cross-validation*) che deriva da un confronto tra le vere classi (tipi di patologie) e le classi previste dalla regola di allocazione. La regola consiste nell'allocare l'unità x^* alla k-esima classe in modo tale da minimizzare il *punteggio discriminante*:

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x^* - \bar{x}'_{ik})^2}{(s_i + s_0)^2} - 2\log \pi_k$$

in cui il primo termine è il quadrato della distanza standardizzata tra centroide e osservazione, mentre il secondo termine è una correzione basata sulle probabilità a priori di ciascuna classe, che stimiamo attraverso la formula:

$$\hat{\pi}_k = \frac{1}{K} .$$

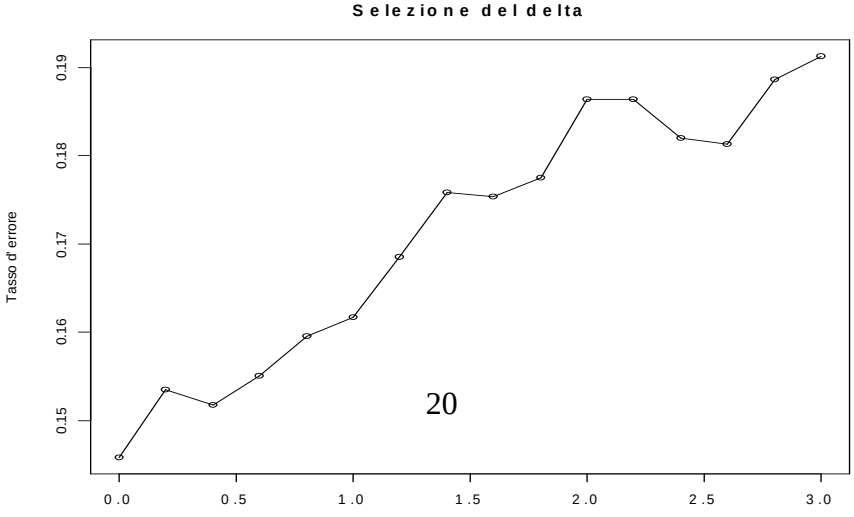


Figura 2.1 Efficacia dell'algoritmo nel *dataset* per diversi valori del parametro Δ

La figura 2.1 mostra la prestazione dell'algoritmo al variare del parametro Δ. Da questa si può notare che il minimo errore si ottiene ponendo $\Delta = 0$, ma questo non risulterebbe utile in quanto manterrebbe lo stesso numero di variabili. E' utile notare che l'errore cresce in maniera abbastanza regolare al crescere di Δ e non supera il 20% . Dato che si ha la necessità di ridurre la dimensionalità del *dataset*, è stato scelto il valore $\Delta = 2.6$ che genere un errore pari a circa 0.18

ALL	Geni comuni	28 30 31 33 187 215 224 248 266 281 311 316 318 326 351
	Geni selezionati nel dataset composto dalle prime 6000 variabili	28 30 31 33 187 215 224 248 266 281 311 316 318 326 351 398 465 479 557 586 678 698 701 704 732 744 750 967 978 986 987 993 1036 1040 1042 1044 1063 1119 1156 1185 1236 1238 1239 1240 1249 1260 1316 1352 1365 1449 1461 1489 1494 1618 1658 1673 1679 1737 1739 1743 1788 1832 1973 2011 2018 2028 2075 2156 2196 2226 2245 2261 2301 2324 2401 2436 2532 2553 2566 2592 2628 2640 2656 2759 2766 2776 2791 2891 2900 2912 2933 2955 3000 3021 3103 3277 3309 3313 3316 3317 3363 3373 3399 3414 3462 3479 3492 3498 3540 3559 3634 3675 3684 3768 3804 3817 3840 3878 3879 3892 3893 3896 3899 4083 4084 4087 4103 4107 4197 4210 4216 4254 4285 4286 4293 4315 4321 4327 4341 4344 4345 4347 4368 4383 4384 4397 4401 4410 4413 4449 4469 4474 4484 4502 4518 4519 4526 4531 4532 4539 4555 4579 4583 4602 4614 4617 4621 4660 4666 4676 4717 4740 4745 4748 4749 4755 4780 4782 4788 4836 4854 4881 4894 4898 4932 4942 4947 4948 5011 5071 5078 5092 5093 5115 5134 5148 5161 5220 5236 5245 5257 5258 5265 5292 5305 5347 5360 5370 5371 5376 5409 5425 5460 5464 5474 5494 5499 5517 5528 5568 5569 5601 5602 5608 5627 5750 5830 5833 5834 5886 5905 5913
	Geni selezionati nel dataset composto dalle ultime 6582 variabili	48 63 67 89 97 107 200 278 301 332 337 386 392 404 413 414 420 474 498 560 563 582 636 648 668 669 720 754 799 802 803 809 823 848 859 862 869 873 940 979 1038 1064 1086 1106 1118 1131 1133 1137 1256 1270 1273 1277 1280 1286 1295 1298 1299 1300 1305 1319 1336 1347 1354 1378 1471 1516 1523 1567 1568 1585 1592 1597 1598 1625 1657 1665 1666 1745 1754 1760 1800 1811 1816 1821 1828 1835 1935 1941 1961 1975 1983 2048 2065 2091 2105 2107 2126 2131 2142 2150 2165 2166 2175 2181 2212 2218 2224 2246 2255 2256 2269 2277 2287 2292 2315 2353 2394 2417 2423 2428 2445 2459 2468 2480 2515 2536 2547 2567 2590 2597 2601 2615 2625 2644 2647 2648 2664 2665 2687 2691 2697 2733 2735 2739 2740 2749 2771 2792 2800 2804 2809 2810 2825 2849 2869 2896 2907 2937 2943 2944 2993 3005 3006 3007 3049 3051 3072 3075 3076 3083 3102 3108 3112 3145 3148 3156 3239 3300 3335 3342 3367 3402 3452 3586 3604 3620 3656 3668 3741 3765 3832 3864 3882 3919 3929 3947 3959 3985 3999 4008 4038 4072 4075 4098 4115 4120 4170 4201 4215 4247 4259 4281 4306 4318 4337 4342 4369 4411 4419 4447 4454 4457 4500 4514 4530 4544 4552 4622 4632 4655 4665 4670 4673 4674 4720 4721 4797 4810 4888 4896 4901 4928 4963 4998 5020 5110 5229 5282 5297 5321 5325 5366 5377 5417 5565 5603 5607 5674 5718 5751 5800 5864 5873 5935 5956 6026 6075 6098 6132 6135 6234 6270 6271 6342 6365 6391 6392 6404 6415 6418 6430 6457 6490 6560
MLL	Geni selezionati nel dataset composto dalle prime 6000 variabili	427 1176 1316 2247 2746 2770 2992 3021 3084 3195 3712 3726 3821 3993 4341 4349 4589 4835 4866 5031 5083 5577 5591 5801 5841 5881 5905 5917 5961
	Geni selezionati nel dataset composto dalle ultime 6582 variabili	86 518 565 615 669 684 809 830 955 1049 1106 1131 1133 1136 1155 1170 1195 1260 1311 1347 1354 1368 1441 1480 1516 1661 1754 1789 1830 1842 1903 1930 2050 2098 2187 2212 2245 2305 2347 2428 2443 2518 2585 2597 2733 2919 2947 3063 3075 3139 3161 3401 3537 4033 4105 4274 4632 4657 4797 4884 4895 5282 5286 5297 5355 5368 5643 5675 5742 5831 5924
AML	Geni comuni	60 560 803 967 969 1201 1352 1358 1407 1567 1816 1992 2018 2156 2176 2445 2518 2580 2735 2814 2842 3007 3242 3322 3658 3882 3896 4083 4170 4384 4428 4625 4962 5266 5399
	Geni selezionati nel dataset composto dalle prime 6000 variabili	5 28 29 30 31 32 33 34 36 73 128 149 172 224 235 248 254 258 264 265 266 270 281 288 311 316 326 351 383 398 416 443 448 451 465 527 557 566 571 572 575 586 624 628 678 698 701 702 704 732 735 744 750 768 780 798 814 821 950 978 986 987 993 1011 1014 1036 1040 1042 1043 1044 1063 1117 1119 1131 1134 1166 1176 1181 1182 1185 1186 1232 1236 1238 1239 1240 1249 1260 1316 1325 1345 1449 1461 1463 1470 1489 1494 1503 1512 1563 1579 1618 1658 1667 1673 1677 1679 1737 1742 1743 1788 1825 1836 1881 1970 1972 1973 1989 2028 2062 2113 2174 2223 2226 2261 2324 2325 2401 2436 2455 2456 2475 2491 2497 2502 2509 2529 2533 2553 2564 2566 2571 2575 2592 2640 2708 2758 2759 2766 2773 2788 2795 2835 2838 2858 2891 2899 2900 2912 2931 2933 2955 2980 3000 3019 3021 3022 3044 3054 3109 3120 3121 3150 3177 3192 3201 3235 3248 3268 3272 3277 3296 3313 3316 3325 3363 3385 3399 3400 3414 3420 3438 3448 3464 3479 3480 3492 3493 3498 3540 3559 3602 3634 3639 3643 3646 3671 3675 3684 3687 3715 3722 3766 3768 3805 3806 3825 3837 3840 3867 3879 3892 3893 3897 3899 3933 3973 4001 4059 4103 4190 4195 4197 4210 4216 4234 4236 4245 4254 4278 4280 4282 4285 4286 4293 4304 4321 4339 4341 4344 4345 4347 4368 4383 4397 4401 4410 4413 4474 4484 4486 4518 4519 4526 4531 4532 4535 4545 4555 4577 4579 4583 4589 4621 4629 4660 4666 4681 4688 4693 4695 4707 4713 4717 4736 4740 4745 4748 4749 4755 4774 4782 4788 4794 4801 4811 4821 4862 4873 4880 4898 4903 4910 4927 4930 4932 4943 4947 4948 4981 4982 4991 5002 5011 5028 5031 5036 5071 5078 5088 5089 5093 5097 5113 5115 5124 5139 5147 5148 5192 5193 5202 5218 5220 5223 5227 5245 5250 5258 5292 5306 5337 5339 5347 5360 5369 5370 5376 5390 5409 5414 5415 5425 5437 5464 5474 5494 5498 5499 5502 5517 5528 5568 5573 5577 5580 5591 5601 5602 5608 5627 5657 5666 5674 5733 5737 5750 5770 5785 5797 5801 5813 5830 5833 5834 5854 5855 5856 5868 5884 5910 5913 5915 5917 5924 5945 5948 5955 5963 5997
	Geni selezionati nel dataset composto dalle ultime 6582 variabili	63 67 86 89 97 107 188 200 208 278 301 310 315 319 332 337 359 413 420 474 489 492 498 514 515 518 520 552 565 582 608 615 617 621 648 668 716 720 746 748 752 754 777 778 781 799 802 830 839 841 859 869 884 915 940 955 1029 1030 1031 1048 1064 1079 1086 1098 1103 1118 1147 1153 1155 1170 1195 1200 1242 1243 1248 1250 1251 1256 1273 1277 1280 1286 1295 1298 1299 1300 1305 1312 1313 1315 1319 1332 1336 1347 1354 1384 1401 1402 1405 1439 1441 1471 1516 1523 1537 1541 1568 1585 1592 1605 1623 1625 1651 1661 1665 1668 1732 1760 1762 1789 1800 1803 1811 1821 1828 1832 1835 1842 1872 1923 1930 1935 1941 1949 1960 1961 1962 1969 1983 2007 2020 2021 2032 2048 2057 2091 2093 2098 2101 2107 2111 2131 2142 2150 2155 2160 2165 2166 2169 2173 2175 2181 2187 2202 2212 2218 2221 2222 2224 2244 2246 2254 2255 2256 2260 2264 2268 2272 2277 2279 2287 2292 2304 2305 2310 2313 2331 2335 2344 2348 2353 2359 2389 2394 2404 2417 2418 2420 2428 2443 2459 2468 2478 2480 2486 2492 2515 2527 2536 2544 2546 2547 2567 2585 2601 2613 2625 2644 2647 2648 2657 2660 2664 2665 2677 2685 2687 2691 2692 2697 2700 2704 2739 2740 2770 2771 2791 2794 2799 2800 2803 2804 2825 2839 2849 2855 2869 2874 2896 2910 2914 2937 2943 2944 2956 2993 3005 3006 3034 3049 3072 3076 3083 3084 3102 3108 3116 3123 3126 3147 3191 3210 3239 3249 3271 3287 3289 3300 3303 3335 3342 3367 3401 3402 3417 3433 3454 3478 3537 3557 3586 3620 3629 3645 3656 3668 3707 3739 3741 3748 3775 3818 3832 3864 3919 3929 3947 3949 3958 3959 3960 3970 3985 3999 4008 4029 4038 4070 4072 4075 4115 4120 4134 4187 4215 4235 4250 4318 4337 4369 4386 4408 4411 4419 4420 4421 4435 4447 4450 4454 4457 4464 4468 4471 4475 4500 4514 4530 4544 4552 4560 4581 4622 4645 4650 4655 4657 4662 4673 4676 4706 4716 4720 4780 4810 4819 4851 4856 4878 4884 4888 4895 4896 4901 4928 4935 4946 4958 4963 4983 4987 5045 5049 5110 5165 5207 5219 5229 5234 5242 5247 5248 5269 5270 5277 5282 5286 5297 5321 5322 5325 5326 5345 5355 5397 5403 5417 5515 5526 5532 5533 5603 5615 5630 5643 5675 5739 5751 5800 5812 5831 5862 5864 5866 5873 5885 5940 5956 6000 6002 6007 6011 6012 6026 6039 6075 60986109 6135 6160 6270 6271 6287 6302 6342 6356 6365 6367 6379 6391 6403 6404 6418 6430 6457 6479 6500 6537 6560

Tabella 2.1 Geni selezionati dall'algoritmo degli shrunken centroids differenziati per tipo di malattia e suddivisi in due gruppi di circa 6000 variabili ciascuno.

La tabella 2.1 mostra la i geni selezionati dal metodo *shrunken centroids* per ogni tipo di malattia. Dalla suddetta lista di geni si è condotta un'analisi volta ad eliminare quei geni che sono stati selezionati per tutte e tre le malattie e che, con qualche probabilità, potrebbero indicare caratteristiche proprie di tessuti o cellule affette da questo tipo di patologia. Dall'analisi si sono selezionati dunque un gruppo di 635 geni.

Si è interessati a capire quanto tale selezione può sbagliare. A questo scopo si costruisce la matrice di confusione in cui si confrontano i valori previsti dalla metodo utilizzato e i reali valori della variabile risposta.

Metodo shrunken centroids applicato sul dataset completo		Classe prevista		
		ALL	MLL	AML
Classe reale	ALL	18	5	1
	MLL	5	14	1
	AML	0	1	27
Tasso d'errore: 0.18				

Tabella 2.2 Matrice di confusione relativa all'applicazione del metodo *shruken centroid* con la validazione incrociata *leave-one-out.*.

La matrice di confusione riportata nella Tabella 2.2 evidenzia un tasso di errore relativamente basso considerata la riduzione di dimensionalità proposta: da 12582 geni si passa a 635 con un errore pari a circa il 20%.

2.3 Analisi esplorativa

2.3.1 Le curve di Andrews

Uno strumento sofisticato per la visualizzazione di dati multidimensionali sono i diagrammi di Andrews (*Andrews plot*). Essi sono particolarmente indicati per individuare unità statistiche simili e/o aberranti (*outliers*). Il metodo consiste nel trasformare ogni osservazione in una serie di

Fourier, ovvero nel mappare una osservazione *p*-dimensionale in uno spazio bidimensionale mediante la trasformazione:

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

al variare del parametro *t* tra $[-\pi, \pi]$.

Le caratteristiche dei dati sono preservate dalle curve di Andrews grazie ad alcune caratteristiche di cui gode la serie sopra presentata. Essa infatti:

1. preserva la media.

   Si indichi con $\underline{\bar{x}}$ il vettore delle medie delle *n* osservazioni:

   $$\underline{\bar{x}} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_p)^T$$

   con

   $$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \quad con\ j = 1, ..., p$$

   allora la trasformazione relativa alla media $\bar{x}$, coincide con la media delle trasformazioni corrispondenti alle *n* osservazioni.

   Si ha quindi:

   $$f_x(t) = \frac{\sum_{i=1}^{n} f_{xi}(t)}{n}.$$

2. Preserva le distanze.

   La distanza tra due osservazioni viene preservata nella trasformazione in serie:

   $$\pi \parallel x - y \parallel^2 = \pi \sum_{i=1}^{p} (x_i - y_i)^2 = \parallel f_x(t) - f_y(t) \parallel = \parallel f_x(t) - f_y(t) \parallel_{L_2} = \int_{-\pi}^{\pi} \left[f_x(t) - f_y(t) \right]^2 dt.$$

   In questo modo si ha proporzionalità tra la distanza che separa le due funzioni calcolate nei due punti *x* e *y* e la distanza Euclidea tra gli stessi punti.

3. Preserva l' ordinamento.

   Se un punto *y* è collocato sulla linea che unisce *x* e *z*, allora per qualsiasi valore di *t*, la $f_y(t)$ è posizionata tra $f_x(t)$ e $f_z(t)$.

4. Preserva la varianza.

Se le osservazioni sono incorrelate e con varianza σ^2, allora la varianza della funzione in *t* è espressa da:

$$\mathrm{var}\left[f_x(t)\right] = \sigma^2\left(\frac{1}{2} + \sin^2 t + \cos^2 t + \sin^2 2t + \cos^2 2t ...\right)$$

Si possono distinguere ora due casi:

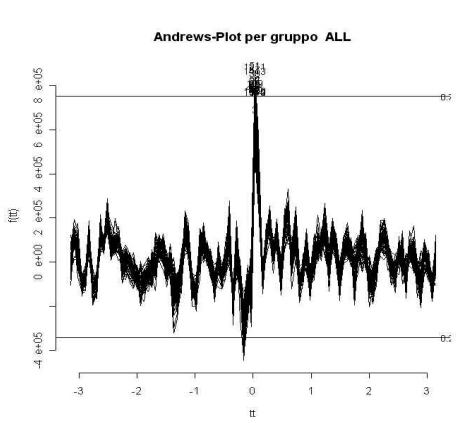
- *p* pari:

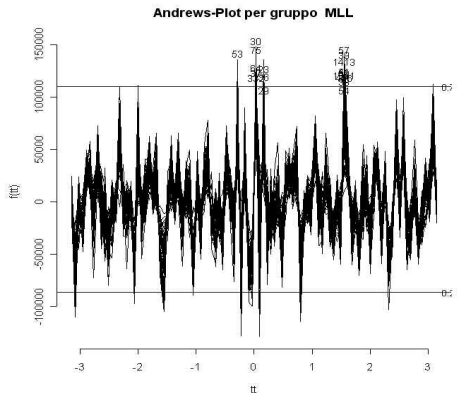
  la varianza si riduce ad una costante, $\frac{1}{2}\sigma^2 p$;

- *p* dispari:

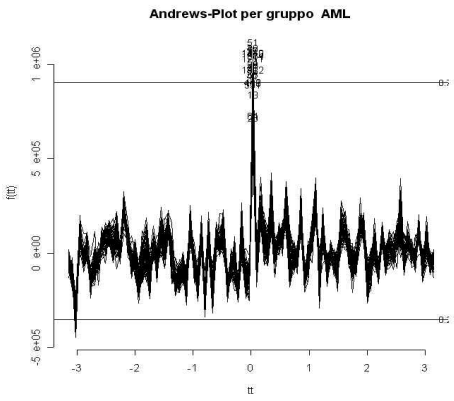
  la varianza varia tra $\sigma^2(p-1)$ e $\sigma^2(p+1)$.

Si noti che nel primo la varianza non dipende da *t* e nel secondo, l'influenza di *t* decresce all'aumentare di *p*. In questo modo la variabilità della funzione è costante su tutto l'intervallo di *t*, cosa che facilita l'interpretazione del grafica.





(a)



(c)

Figura 2.3 (a)Curve di Andrews per la leucemia di tipo ALL nel dataset composto dai geni selezionati dal metodo *shrunken centroid*. (b)Curve di Andrews per la leucemia di tipo MLL nel dataset composto dai geni selezionati dal

metodo *shrunken centroid*; (c)Curve di Andrews per la leucemia di tipo AML nel dataset composto dai geni selezionati dal metodo *shrunken centroid.*

Si nota che le tre tipologie di leucemia sono caratterizzate da diversi profili genetici: l'effettiva differenza sta nel campo di variazione dei tre gruppi; bisogna comunque tenere presente che la forma delle curve di Andrews dipende essenzialmente dal numero di geni di cui si dispone. Dal momento che tale numerosità rimane costante per i tre tipi di patologie la discrepanza tra i tre grafici può essere imputata esclusivamente all'effettiva diversità tra gruppi.
La funzione `andrews.plot`, con l'ausilio della quale si sono tracciati i profili in Figura 2.3, fornisce in output l'elenco dei geni che oltrepassano due bande in modo da lasciare uscire il 25% dei profili sia verso l'alto che verso il basso.

In tabella 2.4, sono evidenziati nei profili di Andrews tutti i geni che sono sovraespressi e sottoespressi. Dall'analisi di questa tabella , si può notare che alcune espressioni genetiche rilevate al di fuori delle bande sono comuni nelle tre patologie, ed i soggetti in cui sono state rilevate tali alterazioni si ripetono. Questo può significare che tali soggetti sono caratterizzati da alterazioni a livello genetico che non dipendono dalla patologia in studio. Un'analisi più approfondita richiederebbe di analizzare tutte le permutazioni possibili delle unità prese tre a tre da ciascun gruppo. In questo elaborato tale analisi verrà tralasciata in quanto richiederebbe tempi di elaborazione troppo lunghi.

Geni le cui curve di Andrews fuoriescono dalle bande per ogni classe	
Livello 1%	
ALL	51 1413
MLL	30 53 75
AML	51 52
Livello 2%	
ALL	45 51 1211 1413
MLL	30 53 75
AML	39 51 52
Livello 3%	
ALL	29 45 51 57 1211 1413
MLL	30 53 75 39 57
AML	39 42 51 52
Livello 4%	
ALL	29 45 51 57 1211 1413
MLL	30 53 75 39 57
AML	39 42 51 52
Livello 5%	
ALL	29 45 51 57 1211 1413
MLL	30 53 75 39 57
AML	39 42 51 52 57
Livello 6%	
ALL	29 39 45 51 52 57 1211 1413
MLL	30 53 75 39 49 57
AML	39 42 51 52 57 1110 1413
Livello 8%	
ALL	29 39 45 51 52 57 1211 1413
MLL	30 53 75 39 49 57 1413
AML	39 42 51 52 57 1110 1211 1413
Livello 10%	
ALL	29 39 45 51 52 54 57 87 109 1211 1312
MLL	23 30 53 64 75 39 45 49 57 98 1413
AML	39 42 51 52 531 57 98 109 1110 1211 1413 1514
Livello 12%	
ALL	29 39 41 45 51 52 54 57 87 98 109 1211 1312 1413
MLL	23 30 53 64 75 39 45 49 51 57 98 1413
AML	13 39 42 51 52 531 57 98 109 1110 1211 1413 1514

Tabella 2.4

Geni le cui curve di Andrews fuoriescono dalle bande per ogni classe	
Livello 14%	
ALL	29 39 41 45 47 51 52 54 57 87 98 109 1211 1312 1413
MLL	23 30 53 64 75 39 41 45 49 51 57 98 1413
AML	13 39 42 51 52 531 57 98 109 1110 1211 1312 1413 1514
Livello 15%	
ALL	29 39 41 45 47 51 52 54 57 87 98 109 1211 1312 1413
MLL	23 30 37 53 64 75 38 39 41 45 49 51 57 98 1413
AML	13 39 42 51 52 531 57 98 109 1110 1211 1312 1413 1514
Livello 16%	
ALL	29 39 41 42 45 47 51 52 54 57 87 98 109 1211 1312 1413
MLL	23 30 37 53 64 75 38 39 41 45 49 51 57 98 1413
AML	13 29 39 42 446 51 52 531 54 57 98 109 1110 1211 1312 1413 1514
Livello 18%	
ALL	29 39 41 42 45 47 49 51 52 54 57 87 98 109 1211 1312 1413
MLL	23 30 37 53 64 75 38 39 41 45 49 51 57 98 1413
AML	13 29 39 41 42 446 48 51 52 531 54 57 98 109 1110 1211 1312 1413 1514
Livello 20%	
ALL	29 39 41 42 446 45 47 49 51 52 54 57 87 98 109 1110 1211 1312 1413
MLL	23 30 36 37 53 64 75 38 39 41 45 49 51 57 98 1211 1413
AML	13 29 38 39 41 42 446 48 51 52 531 54 57 98 109 1110 1211 1312 1413 1514
Livello 22%	
29 39 41 42 446 45 47 49 51 52 54 57 87 98 109 1110 1211 1312 1413 1514	
23 30 36 37 53 64 75 38 39 41 446 45 49 51 54 57 98 1211 1413	
13 20 29 38 39 41 42 446 48 51 52 531 54 57 98 109 1110 1211 1312 1413 1514	

Tabella 2.5

	Livello 24%
ALL	3 29 39 40 41 42 446 45 47 49 51 52 54 57 87 98 109 1110 1211 1312 1413 1514
MLL	23 29 30 332 36 37 53 64 75 38 39 41 446 45 49 51 54 57 98 1211 1413
AML	13 20 29 64 38 39 41 42 446 47 48 51 52 531 54 55 57 98 109 1110 1211 1312 1413 1514
	Livello 25%
ALL	3 29 39 40 41 42 446 45 47 49 51 52 54 57 87 98 109 1110 1211 1312 1413 1514
MLL	23 29 30 332 36 37 53 64 75 38 39 41 446 45 49 51 54 57 98 1211 1413
AML	13 20 29 64 38 39 41 42 446 47 48 51 52 531 54 55 57 98 109 1110 1211 1312 1413 1514
	Livello 26%
ALL	3 27 29 39 40 41 42 446 45 47 49 51 52 54 55 57 87 98 109 1110 1211 1312 1413 1514
MLL	23 29 30 332 36 37 53 64 75 38 39 41 446 45 49 51 54 57 87 98 1211 1413
AML	13 20 29 64 38 39 41 42 446 47 48 51 52 531 54 55 57 87 98 109 1110 1211 1312 1413 1514
	Livello 28%
ALL	3 27 29 38 39 40 41 42 446 45 47 49 51 52 54 55 57 87 98 109 1110 1211 1312 1413 1514
MLL	23 29 30 32 332 36 37 53 64 75 38 39 41 446 45 49 51 54 57 87 98 1211 1413
AML	13 20 29 64 38 39 41 42 446 47 48 51 52 531 54 55 56 57 87 98 109 1110 1211 1312 1413 1514
	Livello 30%
ALL	3 27 29 38 39 40 41 42 43 446 45 47 48 49 51 52 54 55 57 87 98 109 1110 1211 1312 1413 1514
MLL	23 25 27 29 30 32 332 36 37 53 64 75 38 39 41 446 45 49 51 52 54 57 87 98 1211 1413 1514
AML	3 5 13 20 27 29 64 38 39 41 42 43 446 47 48 51 52 531 54 55 56 57 87 98 109 1110 1211 1312 1413 1514

Tabella 2.6

Tabella 2.4 e Tabella 2.5 e Tabella 2.7 Geni che generano profili di Andrews che fuoriescono per ogni gruppi dalle bande fissate a vari livelli.

Per ciascun gruppo è necessario mettere a confronto le espressioni genetiche che fuoriescono dalle bande per avere una prima idea su quali di essi le curve hanno comportamenti diversi a seconda della patologia. E' utile inoltre estrapolare più informazioni possibili riguardanti i geni anomali per ogni patologia presa singolarmente: è infatti possibile che un gene abbia un buon potere discriminante se il suo profilo di Andrews fuoriesce (o non fuoriesce) dalle bande per un solo tipo i patologia. Si può notare da una prima analisi quindi che non esistono grosse differenze

tra patologie: in linea di massima i geni che presentano curve di Andrews con un campo di variazione più ampio in un gruppo, presentano una variazione anomala anche negli alti gruppi, e di conseguenza, i geni che hanno andamento regolare in gruppo godono della stessa caratteristica negli alti. E' probabile che tale anomalia sia dovuta ad una alterazione dei geni provocata da tutte e tre le forme di leucemia o, in alternativa, sia dovuta ad altri tipi di patologie direttamente collegate con la leucemia dipendenti dal fatto che la malattia rende le difese immunitarie di ogni paziente meno potenti.

2.3.2 Analisi di ragruppamento sui geni

Mentre l'analisi grafica condotta attraverso l'analisi delle curve di Andrews conduce a considerazioni sulla variazioni dei livelli di espressione dei geni da patologia a patologia sottolineando la presenza di alterazioni genetiche non dipendenti dalla patologia, l'analisi di raggruppamento si occupa 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 ed associati a gruppi diversi.

L'idea alla base di tale studio è di fa emergere all'interno del *dataset* gruppi di unità affini od omogenei. Il concetto di affinità o somiglianza viene tradotto nel linguaggio statistico con la nozione di *distanza*: quanto più due osservazioni sono vicine tanto più sono *simili* tra loro.

	VARIABILI NON STANDARDIZZATE		Suddivisione in tre classi			Suddivisione in quattro classi			
	METODO	PATOLOGIA	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
METODI BASATI SULLE PARTIZIONI	METODO DEI MEDOIDI	ALL	225	393	17	37	393	14	191
		MLL	20	242	373	34	212	16	373
		AML	19	194	492	17	154	67	397
	METODO DELLE K-MEDIE	ALL	44	17	577	171	14	35	415
		MLL	45	18	572	419	175	25	16
		AML	82	16	587	494	16	51	74
METODI GERARCHICI	METODO DEL LEGAME SINGLOLO	ALL	633	1	1	632	1	1	1
		MLL	1	633	1	1	630	1	3
		AML	1	633	1	632	1	1	1

	METODO DEL LEGAME MEDIO	ALL	619	14	2	619	14	1	1
		MLL	2	618	15	2	618	11	4
		AML	15	619	1	2	619	13	1
	METODO DEL LEGAME COMPLETO	ALL	605	14	16	605	11	16	3
		MLL	35	592	8	12	592	23	8
		AML	17	611	7	9	611	8	7

Tabella 2.7 Numerosità dei cluster per le diverse tecniche di raggruppamento applicate a 635 geni della selezione avvenuta attraverso il metodo "nearest shrunken centroids" utilizzando variabili non standardizzate. La parte sinistra propone la suddivisione in tre gruppi, quella a destra in quattro.

	VARIABILI STANDARDIZZATE		Suddivisione in tre classi			Suddivisione in quattro classi			
	METODO	PATOLOGIA	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
METODI BASATI SULLE PARTIZIONI	METODO DEI MEDOIDI	ALL	226	392	17	37	392	14	192
		MLL	21	242	372	17	213	33	372
		AML	18	196	421	16	156	68	395
	METODO DELLE K-MEDIE	ALL	34	19	382	41	34	541	19
		MLL	116	38	481	131	31	16	457
		AML	82	16	587	512	68	40	15
METODI GERARCHICI	METODO DEL LEGAME SINGLOLO	ALL	623	11	1	623	7	4	1
		MLL	631	3	1	629	3	2	1
		AML	625	7	3	625	5	3	2
	METODO DEL LEGAME MEDIO	ALL	608	26	1	608	11	15	1
		MLL	565	51	19	565	51	4	15
		AML	15	613	7	6	613	7	9
	METODO DEL LEGAME COMPLETO	ALL	592	12	31	592	11	31	1
		MLL	584	32	19	584	32	4	15
		AML	12	613	10	12	534	79	10

Tabella 2.8 Numerosità dei cluster per le diverse tecniche di raggruppamento applicate a 635 geni della selezione avvenuta attraverso il metodo "nearest shrunken centroid" utilizzando variabili standardizzate. La parte sinistra propone la suddivisione in tre gruppi, quella a destra in quattro.

Per misurare la distanza tra cluster si sono utilizzati da principio due metodi di partizione:

a. metodo delle *k*-medie (*k-means*) introdotto nel 1967 da Mc Queen rif [39]. Tale metodo consiste in un algoritmo composto da 4 passi:

1. suddivisione casuale delle osservazioni in *g*;
2. per i *g* gruppi si calcolino i centroide (le medie aritmetiche);
3. allocazione di ciascuna osservazione al centroide più vicino;
4. ritorno al passo due. L'algoritmo termina quando le osservazioni non si spostano più da un cluster all'altro.

Tale metodo viene utilizzato generalmente per dati continui in cui si calcola la distanza euclidea.

b. Metodo dei medoidi introdotto nel 1987 da Kaufman e Rousseeuw rif [40]. Tale algoritmo utilizza lo stesso algoritmo proposto per le *k*-medie, ma, invece di basarsi sui centroidi, si basa sui medoidi (una delle osservazioni posizionata al centro del cluster).

Si sono poi riproposte le analisi utilizzando metodi gerarchici che prevedono algoritmi di tipo agglomerativo, e che sono caratterizzati dalla possibilità di suddividere le osservazioni senza avere il numero di gruppi previsti a priori. Per questo tipo di analisi si necessita di definire una misura di distanza tra cluster. Quelle da noi utilizzate sono:

1. legame singolo (*single linkage*) in cui la distanza tra due cluster, C_1 e C_2, viene definita come:

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

dove *d(i,j)* rappresenta la distanza tra l'*i*-esimo e il *j*-esimo gene.

2. Legame completo (*complete linkage*), in cui la distanza tra due cluster, C_1 e C_2, viene definita come:

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

3. Legame medio (*average linkage*), in cui la distanza tra due cluster, C_1 e C_2, viene definita come:

$$D(C_1, C_2) = \frac{\sum d(i, j)}{n_{C1} + n_{C2}} \quad i \in C_1, \ j \in C_2.$$

Applicando al dataset composto dalle variabili selezionate dall'algoritmo di Tibshirani le varie metodologie per la formazione di 3 cluster si può notare che se ne forma uno di grandi dimensioni e gli altri di dimensioni più piccole, passando da 3 a 4 cluster si nota che i gruppi che cambiano dimensione sono quelli composti da meno variabili. Confrontando infine i vari metodi si può notare i metodi da cui si ottengono *cluster* con numerosità più equilibrata sono i metodi gerarchici che utilizzano legame medio o legame completo. Inoltre nel passaggio da quattro a tre gruppi, il cluster più corposo mantiene intatta la sua dimensione, ossia vengono aggregati i *cluster* meno numerosi.

Si è poi passati all'utilizzo dei metodi non gerarchici di raggruppamento basati sulla partizione dello spazio. Sebbene tali metodi prevedano la conoscenza a priori del numero di gruppi in cui vengono suddivise le unità, si è deciso, in base ai risultati ottenuti dai metodi gerarchici di utilizzare il metodo delle *k*-medie e dei medoidi per la suddivisione dei geni in tre o quattro gruppi. Anche questi risultati sono riassunti nelle Tabelle 2.7 e 2.8 e confermano l'analisi precedente: sembra dunque sensato suddividere le espressioni genetiche in tre o quattro gruppi in quanto anche in questo caso viene a formarsi un gruppo piuttosto consistente rispetto agli alti.

Geni selezionati come medoidi (variabili non standardizzate)							
Patologia	Analisi con tre gruppi			Analisi con quattro gruppi			
ALL	3639	4998	702	1742	4998	702	3639
MLL	984	3897	5247	1742	3897	984	5247
AML	254	2758	4998	987	4998	1166	1401

Tabella 2.9

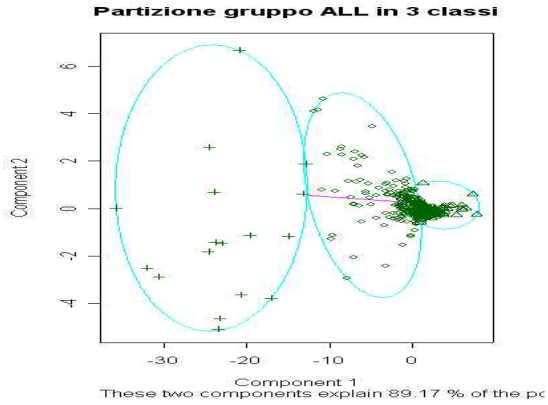
Geni selezionati come medoidi (variabili standardizzate)							
Patologia	Analisi con tre gruppi			Analisi con quattro gruppi			
ALL	3639	4998	702	1742	4998	702	3639
MLL	702	3897	5247	984	3897	1742	5247
AML	984	2758	4998	984	4982	1166	1401

Tabella 2.10

Tabella 2.9 e Tabella 2.10 Tabelle riassuntive dei geni selezionati come medoidi per le tre forme di leucemia facendo uso di variabili standardizzate (Tabella 2.10) e non (Tabella 2.9). La parte 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.

Le analisi proposte nelle Tabelle 2.9 e 2.10 dimostrano che alcuni geni selezionati come medoidi nelle variabili non standardizzate hanno perso tale caratteristica passando a variabili standardizzate ma c'e da sottolineate che gran parte dei medoidi risultano comuni alle due analisi.

Dato che il metodo dei medoidi è basato sulla ripartizione dello spazio si forniscono di seguito i grafici che rappresentano esattamente i raggruppamenti in tre o quattro *cluster*:



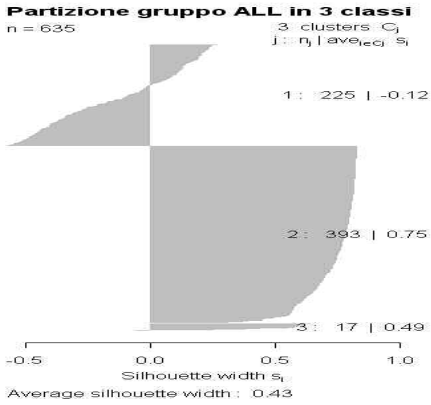
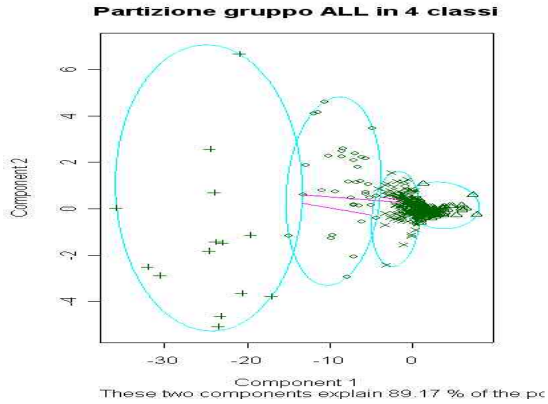


Grafico 2.11

Grafico 2.12



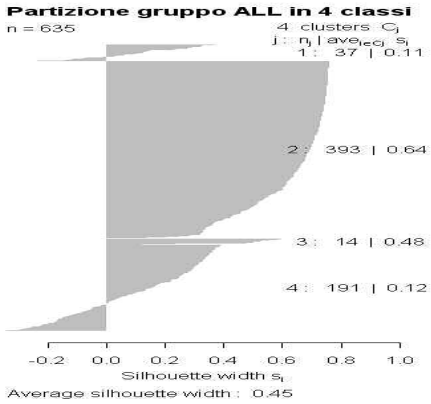
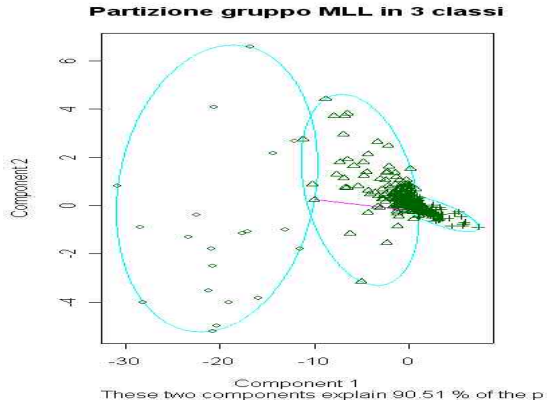


Grafico 2.13

Grafico 2.14

Grafico 2.11 e Grafico 2.12 e Grafico 2.13 e Grafico 2.14 Indicano la partizione dello spazio tramite il metodo dei medoidi in tre (Grafici 2.11 e 2.12) e quattro (Grafici 2.13 e 2.14) classi per la patologia di tipo ALL su dati non standardizzati.



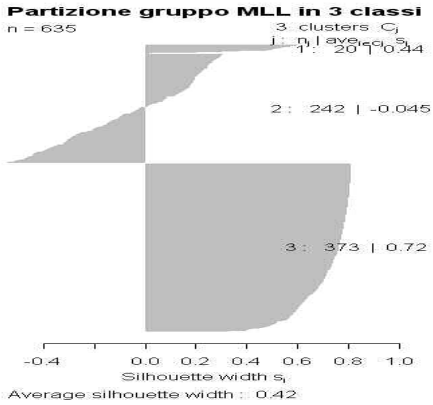
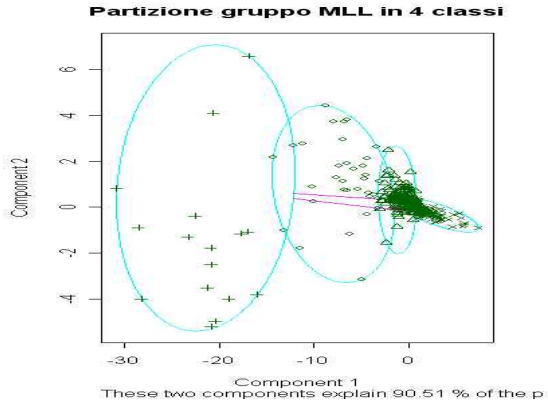


Grafico 2.15

Grafico 2.16



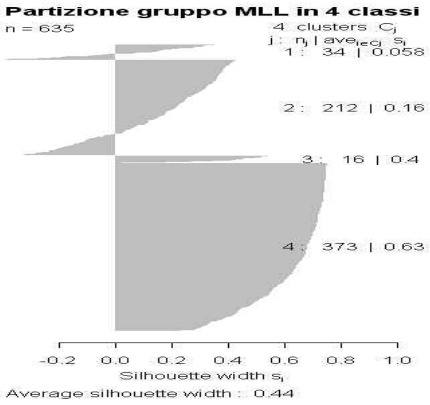
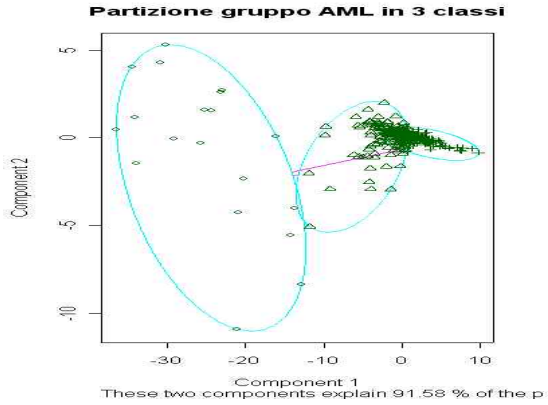


Grafico 2.17

Grafico 2.18

Grafico 2.15 e Grafico 2.16 e Grafico 2.17 e Grafico 2.18 Indicano la partizione dello spazio tramite il metodo dei medoidi in tre (Grafici 2.15 e 2.16) e quattro (Grafici 2.17 e 2.18) classi per la patologia di tipo MLL su dati non standardizzati.



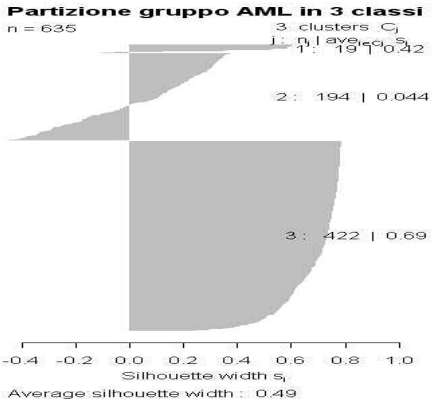
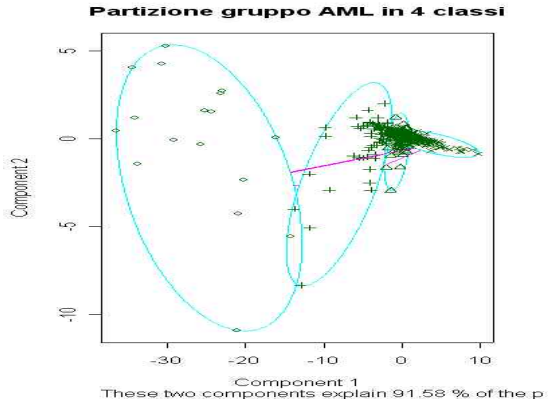


Grafico 2.19

Grafico 2.20



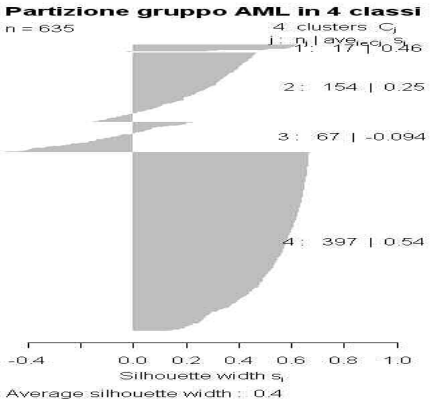


Grafico 2.21

Grafico 2.22

Grafico 2.15 e Grafico 2.16 e Grafico 2.17 e Grafico 2.18 Indicano la partizione dello spazio tramite il metodo dei medoidi in tre (Grafici 2.15 e 2.16) e quattro (Grafici 2.17 e 2.18) classi per la patologia di tipo MLL su dati non standardizzati.

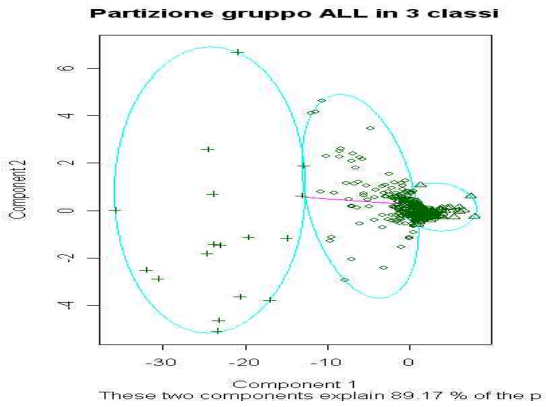


Grafico 2.23

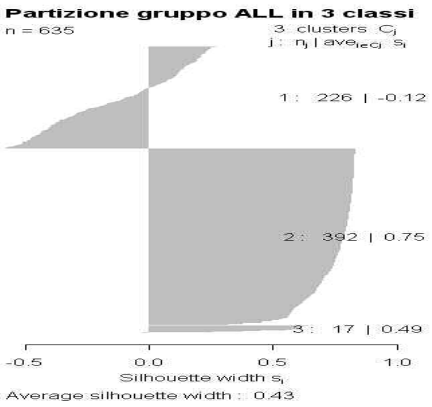


Grafico 2.24

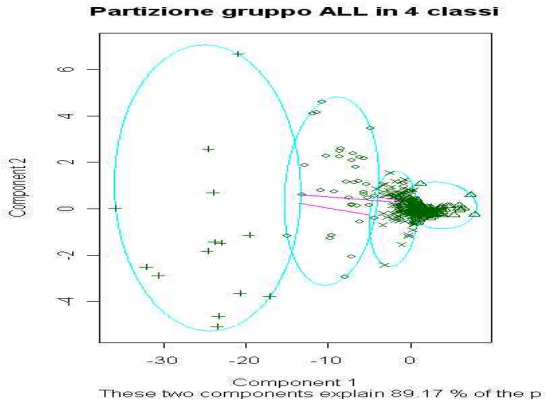


Grafico 2.25

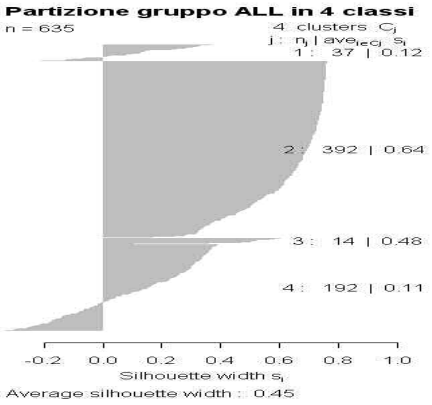


Grafico 2.26

Grafico 2.23 e Grafico 2.24 e Grafico 2.25 e Grafico 2.26 Indicano la partizione dello spazio tramite il metodo dei medoidi in tre (Grafici 2.23 e 2.24) e quattro (Grafici 2.25 e 2.26) classi per la patologia di tipo ALL su dati standardizzati.

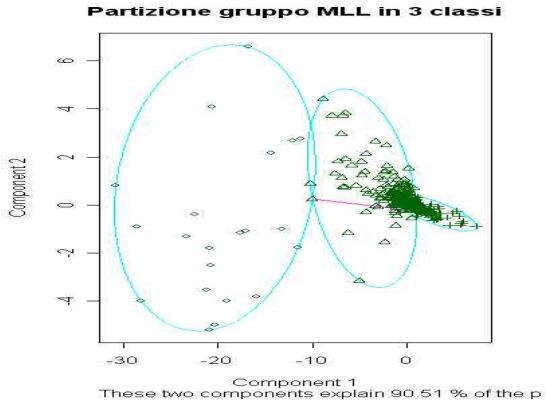


Grafico 2.27

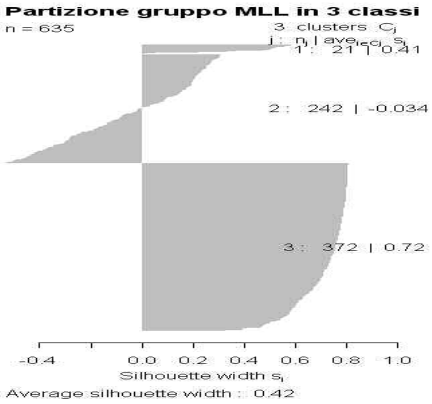


Grafico 2.28

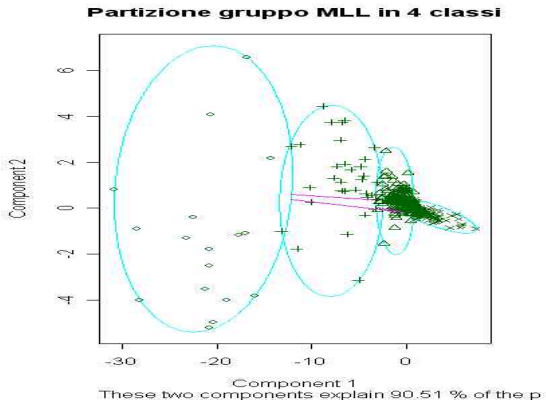


Grafico 2.29

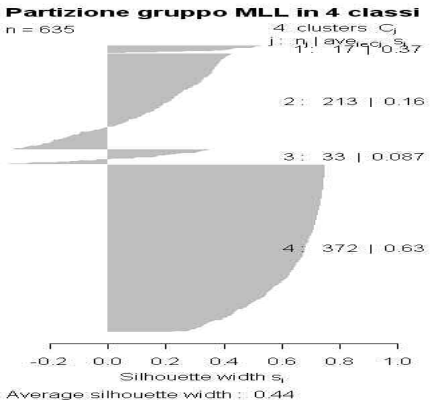


Grafico 2.30

Grafico 2.27 e Grafico 2.28 e Grafico 2.29 e Grafico 2.30 Indicano la partizione dello spazio tramite il metodo dei medoidi in tre (Grafici 2.27 e 2.28) e quattro (Grafici 2.29 e 2.30) classi per la patologia di tipo MLL su dati standardizzati.

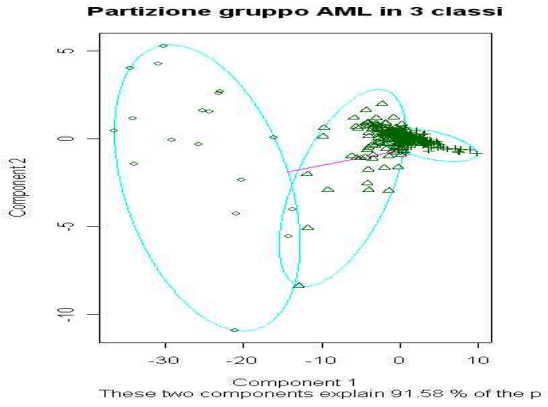


Grafico 2.31

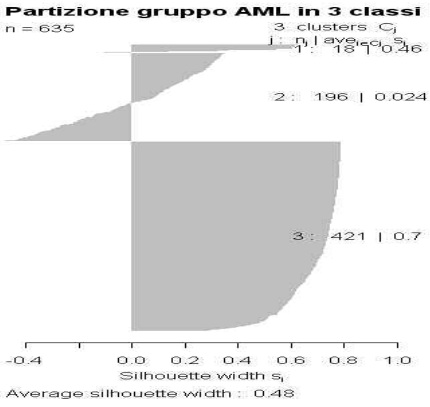


Grafico 2.32

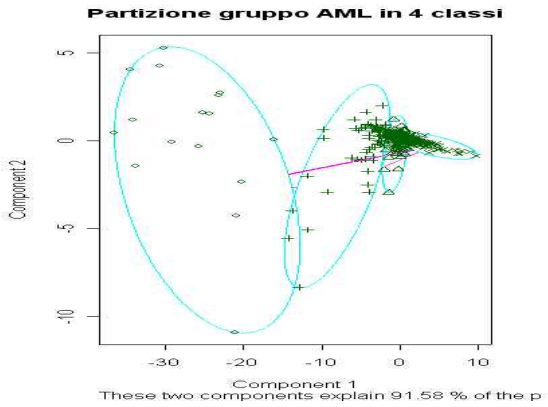


Grafico 2.33

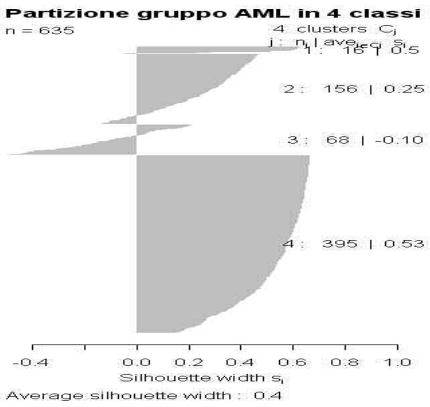


Grafico 2.34

Grafico 2.31 e Grafico 2.32 e Grafico 2.33 e Grafico 2.34 Indicano la partizione dello spazio tramite il metodo dei medoidi in tre (Grafici 2.31 e 2.32) e quattro (Grafici 2.33 e 2.34) classi per la patologia di tipo AML su dati standardizzati.

Per rappresentare in maniera grafica l'analisi di raggruppamento di tipo gerarchico si presentano di seguito dendrogrammi relativi ai legame medio in quanto, i cluster individuati attraverso tale legame presentano numerosità più equilibrate. Tali dendrogrammi sono costruiti a partire dal dataset ridotto tramite l'algoritmo Shrunken Centroids. Inoltre per ciascuna patologia si affiancano i risultati ottenuti dalle diverse applicazioni sia su variabili standardizzate che non.

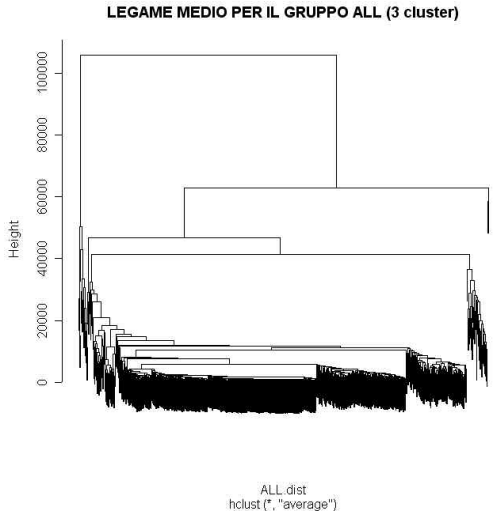


Figura 2.35 Dendrogramma relativo al legame medio relativo al gruppo ALL. Analisi condotta su variabili non standardizzate.

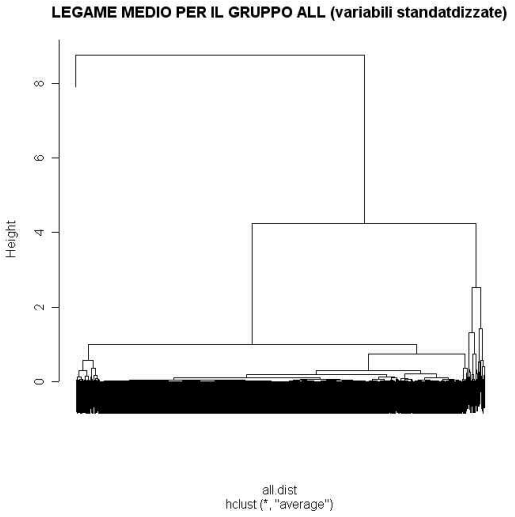


Figura 2.36 Dendrogramma relativo al legame medio relativo al gruppo ALL. Analisi condotta su variabili standardizzate.

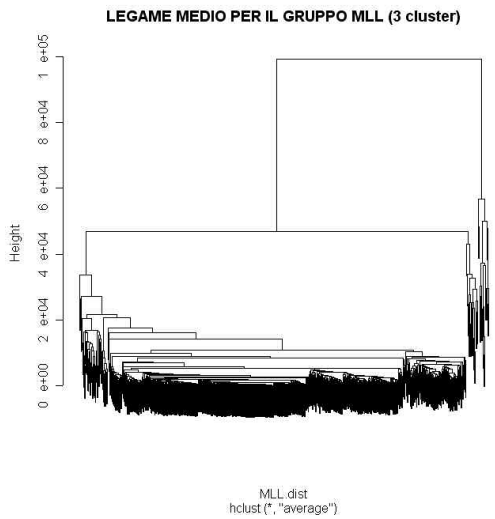


Figura 2.37 Dendrogramma relativo al legame medio relativo al gruppo MLL. Analisi condotta su variabili non standardizzate.

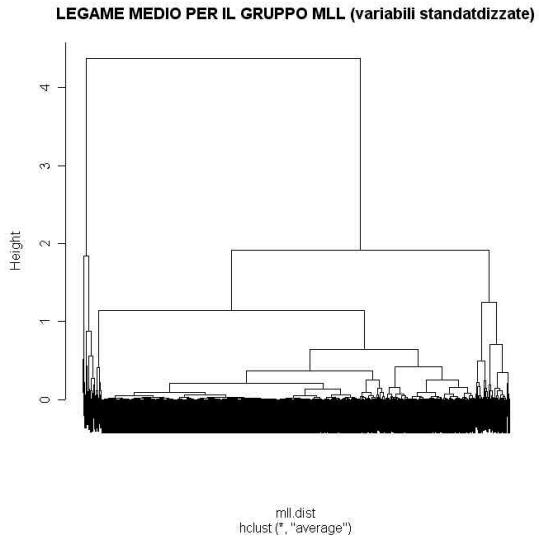


Figura 2.38 Dendrogramma relativo al legame medio relativo al gruppo MLL. Analisi condotta su variabili non standardizzate.

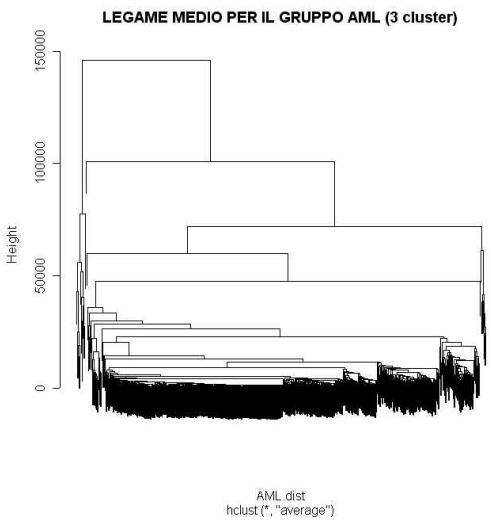


Figura 2.39 Dendrogramma relativo al legame medio relativo al gruppo AML. Analisi condotta su variabili non standardizzate.

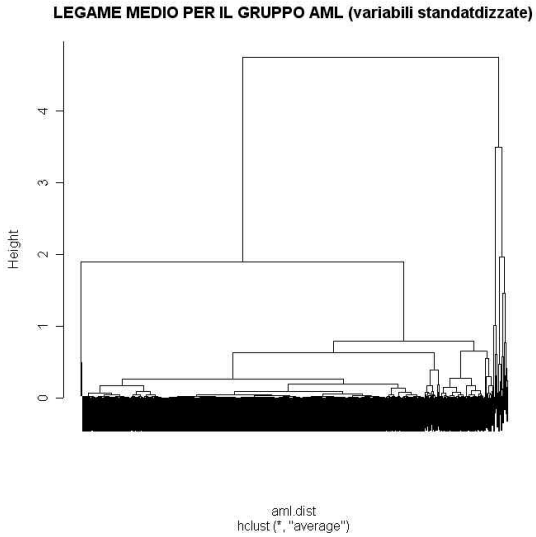


Figura 2.40 Dendrogramma relativo al legame medio relativo al gruppo AML. Analisi condotta su variabili standardizzate.

Dopo tali analisi non risulta ancora possibile delineare conclusioni sensate, per poterlo fare avremmo bisogno di accurate analisi biologiche in grado di rilevare da una parte le funzioni dei geni appartenenti ai diversi *cluster* o, nel caso dell'analisi non gerarchica, dei medoidi; dall'altra fornire delle spiegazioni sul significato biologico che può avere il modo in cui vengono aggregati i geni nei *cluster*; infine motivare la presenza di un *cluster* più numeroso degli altri per ogni gruppo.

APPENDICE CAPITOLO 2

rut.2.1 #caricamento del dataset

```
train<-read.table("C:\\tesi\\train.txt",sep=",",header=F)
test<-read.table("C:\\tesi\\test.txt",sep=",",header=F)
train1<-train[,-12583]
test1<-test[,-12583]
dati<-rbind(train,test)
dati1<-rbind(train1,test1)

ga<-rbind(train[1:20,],test[1:4,])
gb<-rbind(train[21:37,],test[5:7,])
gc<-rbind(train[38:57,],test[8:15,])

ga1<-ga[,-12583]
gb1<-gb[,-12583]
gc1<-gc[,-12583]

dat1<-rbind(ga,gb,gc)
dat2<-dat1[,-12583]
```

rut.2.2 # Selezione dei geni attraverso il metodo dei shrunken centroids

```
sr.centroids<-function(ga,gb,gc,data,prior=c(1/3,1/3,1/3),delta,x)
{

      data<-t(data)
      ga<-t(ga)
      gb<-t(gb)
      gc<-t(gc)
      ma<-apply(ga,1,mean)
      mb<-apply(gb,1,mean)
      mc<-apply(gc,1,mean)
      mg<-apply(data,1,mean)


      ssa<-apply(((ga-ma)**2),1,sum)
      ssb<-apply(((gb-mb)**2),1,sum)
      ssc<-apply(((gc-mc)**2),1,sum)

      si<-sqrt((1/(ncol(data)-3))*(ssa+ssb+ssc))
      so<-median(si)
      m<-rep(0,3)

      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)))

      da<-(ma-mg)/(m[1]*(si+so))
      db<-(mb-mg)/(m[2]*(si+so))
      dc<-(mc-mg)/(m[3]*(si+so))
      p1<-abs(da)-delta
      p2<-abs(db)-delta
      p3<-abs(dc)-delta
      dda<-parte.pos(p1)$y
      ddb<-parte.pos(p2)$y
      ddc<-parte.pos(p3)$y
      dda<-sign(da)*dda
      ddb<-sign(db)*ddb
      ddc<-sign(dc)*ddc
```

```
      xa<-mg+m[1]*(si+so)*dda
      xb<-mg+m[2]*(si+so)*ddb
      xc<-mg+m[3]*(si+so)*ddc
      punt<-rep(0,3)
      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])
      pa<-parte.pos(abs(da)-delta)$pos
      pb<-parte.pos(abs(db)-delta)$pos
      pc<-parte.pos(abs(dc)-delta)$pos
      gr<-which.min(punt)
      list(gruppo=gr,punteggi=punt,genia=pa,genib=pb,genic=pc)
}
```

rut.2.3 `#Funzione che calcola la 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)
}
```

rut.2.4 `#Funzione per per la cross-validation con i valori previsti dal metodo degli schrunken centroids`

```
sr.centroids.cv<-function(ga,gb,gc,data,prior=c(1/3,1/3,1/3),delta){
      vn<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,
      2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,
      3,3,3,3)
      me<-0
      data<-t(data)
      ga<-t(ga)
      gb<-t(gb)
      gc<-t(gc)
      gruppo<-NULL
      for(i in 1:ncol(data)){
      me<-me+1
      x<-data[,me]
      data1<-data[,-me]
      if(me<=24){
                  gga<-ga[,-me]
                  ggb<-gb
                  ggc<-gc
            }
      if((me>24)&&(me<=44)){
                  mme<-me-24
                  ggb<-gb[,-mme]
                  gga<-ga
                  ggc<-gc
            }
```

```
      if(me>44){
                  mme<-me-44
                  ggc<-gc[,-mme]
                  ggb<-gb
                  gga<-ga
            }
            gruppo[i]<-sr.centroids(t(gga),t(ggb),t(ggc),t(data1),prior=prior,
            delta=delta,x)$gruppo
      }
      g<-sr.centroids(t(gga),t(ggb),t(ggc),t(data1),prior=prior,delta=delta,x)
      f<-rep(1,72)
      err<-(72-sum(f[gruppo==vn]))/72
      list(ragg=gruppo,tasso.errore=err,genia=g$genia,genib=g$genib,genic=g$genic)
}
```

rut.2.5 #Funzione per calcolare il tasso di errore per diversi valori di delta

```
rep.srcentr<-function(ga,gb,gc,data,prior=c(1/3,1/3,1/3),delta){
tasso.err<-NULL
for(i in 1:length(delta)){
      tasso.err[i]<-
      sr.centroids.cv(ga,gb,gc,data,prior=prior,delta[i])$tasso.errore
}
tasso.err
}

x<-t(dat2)
```

rut.2.6 #Funzione che seleziona gli elementi uguali per due diversi vettori

```
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
}
```

rut.2.7 #Funzione che seleziona gli elementi diversi per due diversi vettori

```
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
}
```

rut.2.8 #Funzione che standardizza i dati

```
standard<-function(m)
{
        mm <- sweep(m, 2, apply(m, 2, mean))
        sweep(mm, 2, sqrt(apply(m, 2, var)), FUN = "/")
}
```

rut.2.9 #Funzioni per la costruzione delle curve di Andrews

```
trig.fn <- function(x)
        {
                n <- length(x)
                ergebnis <- numeric(n)
                gerade <- (1:n)[(1:n) %% 2 == 0]
                ungerade <- (1:n)[(1:n) %% 2 != 0]
                ergebnis[ungerade] <- sin(x[ungerade])
                ergebnis[gerade] <- cos(x[gerade])
                ergebnis
        }

andrews<-function(x, tt, n.col)
{
        n.tt <- length(tt)
        andrews.x <- rep(x[1]/sqrt(2), n.tt)
        koef <- rep(1:(n.col - 1), rep(2, n.col - 1))[1:(n.col - 1)]
        tt.m<-as.vector(koef)*matrix(tt,nrow=n.col-1,ncol=n.tt,byrow=T)
        tt.m<-apply(tt.m,2,FUN=trig.fn)
        andrews.x<-andrews.x+x[2:n.col]%*%tt.m
#        for(i in 1:n.tt) {
#                andrews.x[i] <- andrews.x[i] +
#                        sum(x[2:n.col] * trig.fn(tt[i] * koef))
#       }
        andrews.x
}

andrews.plot<-
function(m, stand = T,ug=0.25,og=.75,titel=deparse(substitute(m)))
{
        n.row <- dim(m)[1]
        n.col <- dim(m)[2]
        tt <- seq( - pi, pi, length = 100)
        if(!stand)
                m <- standard(m)
        max.andrews <- 0
        min.andrews <- 0
        result <- matrix(0, ncol = length(tt), nrow = n.row)
        for(i in 1:n.row) {
                result[i,  ] <- andrews(m[i,  ], tt, n.col)
        }
#       result<-matrix(apply(m,1,FUN=andrews,tt,n.col),
#                ncol=length(tt),nrow=n.row,byrow=T)
        range.andrews <- range(result)
        plot(0, xlim = c( - pi, pi), ylim = range.andrews,
                axes = F, ylab = "f(tt)",xlab = "tt", type = "n")
```

```
        axis(1)
        axis(2)
        for(i in 1:n.row)
                lines(tt, result[i,  ])
        result.range<-apply(result,1,range)
        range.ug<-quantile(result.range[1,],probs=ug)
        range.og<-quantile(result.range[2,],probs=og)

        vgl<-function(x,y) any(x<y[1] | x>y[2])
        ausserhalb<-apply(result,1,vgl,c(range.ug,range.og))
        abline(h=c(range.ug,range.og))
        text(par()$usr[2],range.ug,ug,cex=.8)
        text(par()$usr[2],range.og,og,cex=.8)
        print("Geni fuori dalle bande:")
        if(length(dimnames(m)[[1]])==0) names.m<-seq(n.row) else
                                names.m<-dimnames(m)[[1]]
        print(names.m[ausserhalb])
        for(i in (1:n.row)[ausserhalb]){
                wo.y<-max(
                        abs(
                        c(max(result[i,]),min(result[i,]))
                        ))
                if(wo.y==abs(min(result[i,])))
                        wo.y<-wo.y*sign(min(result[i,]))
                wo.x<-tt[result[i,]==wo.y]
                text(
                        wo.x,
                        wo.y+sign(wo.y)*0.02*diff(range.andrews),
                        names.m[i],cex=.9
                        )
        }
        title(paste("Andrews-Plot per gruppo ",titel))
        seq(n.row)[ausserhalb]
}
```

Capitolo 3

La regressione logistica penalizzata

3.1 Introduzione

Lo scopo della nostra analisi è quello di individuare una regola di classificazione delle unità statistiche (pazienti) in classi corrispondenti a differenti tipi di patologia, a partire da gruppi di espressioni geniche. Quando le classi, quindi i tipi di patologia, sono solamente due, la regressione logistica può fornire un valido strumento per la modellazione della probabilità di appartenenza ad una classe attraverso la combinazione lineari delle variabili esplicative. Tuttavia, la classica regressione logistica non può essere applicata a dati derivanti da *microarray*, in quanto essi hanno la caratteristica di avere molte più variabili rispetto alle osservazioni. Il primo problema che ne deriva è la multicollinearità, che porta a non avere una soluzione unica e stabile. Il secondo problema è la sovraparametrizzazione che potrebbe infierire nelle capacità previsive del modello .

Per la soluzione di tale problema, la letteratura propone la regressione logistica con verosimiglianza penalizzata in cui non si opera una selezione delle variabili ma vengono considerati tutti i geni con lo stesso peso all'interno del modello.

3.2 La regressione logistica

3.2.1 Presentazione

L' applicazione regressione logistica usuale ai dati di microarray comporta problemi perché il numero di variabili supera in genere il numero di osservazioni. Come si è gia avuto modo di evidenziare nell'introduzione a questo elaborato, la presenza di un numero estremamente elevato di variabili rispetto alle unità statistiche a disposizione fa sorgere essenzialmente due tipi di problemi: la multicollinearità e la sovraparametrizzazione.

Dal primo scaturiscono sistemi di equazioni che non hanno soluzioni uniche e stabili, dal secondo consegue che il modello si adatta molto bene al dataset rispetto al quale è stimato ma non si comporta altrettanto bene se applicato come regola di classificazione di nuove unità.

Il problema della multicollinearità nei dati da microarray è facilmente intuibile e distinguibile in due tipologie: la multicollinearità fondamentale, causata dal fatto che 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 trovare nei processi che portano dal colore al dato numerico.

Il problema della sovraparametrizzazione è causato dalla presenza di così tante variabili in un campione ristretto che consente di modellazioni perfette ma spesso prive di senso.

3.2.2 Formalizzazione

Si considerino da principio due classi: G_1 e G_2. Sia inoltre Y_i la variabile che stabilisce l'appartenenza o meno ad una delle due classi, definita come:

$$Y_i = \begin{cases} 1 & i \in G_1 \\ 0 & i \in G_2 \end{cases}$$

sia inoltre:

$$\Pr[Y_i = 1] = p_i \qquad \Pr[Y_i = 0] = 1 - p_i$$

Lo scopo è quello di individuare come la probabilità di appartenenza dell' *i*-esimo individuo alla classe G_j dipenda dai livelli di espressioni geniche x_i. Una possibile idea potrebbe essere di considerare il modello:

$$p_i = \alpha + \beta x_i \quad i = 1,...,n$$

e quindi stimare i parametri α e β attraverso il semplice modello regressione lineare standard. Tuttavia, tale soluzione non garantisce l'appartenenza all'intervallo [0,1] di p_i, caratteristica base delle probabilità.

La soluzione è quella di trasformare p_i in η_i nel modo seguente:

$$\eta = \log \frac{p_i}{1 - p_i} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \quad i = 1,...,n\,,$$

da cui si ottiene

$$p_i = \frac{\exp(\underline{\beta}^T \underline{x}_i)}{1 + \exp(\underline{\beta}^T \underline{x}_i)} = p(\underline{x}_i, \underline{\beta})$$

$$1 - p_i = \frac{1}{1 + \exp(\underline{\beta}^T \underline{x}_i)} = 1 - p(\underline{x}_i, \underline{\beta})$$

con $\underline{\beta}^T = (\alpha_1, \beta_1, \ldots \beta_p)$ e $\underline{x}_i^T = (1, x_{i1}, ..., x_{ip})$

Una regola pratica in questo tipo di analisi ci dice che ad un modesto numero di variabili esplicative (meno di 15) dovrebbe corrispondere un numero di osservazioni almeno 5 volte più grande(più di 45 osservazioni). Nel caso dei microarray siamo nel caso in cui il numero di variabili è molto superiore al numero di osservazioni. Ci troviamo quindi in una situazione capovolta rispetto alle casistiche usuali.

3.2.3 La penalizzazione nella regressione logistica

Per comprendere il funzionamento della penalizzazione, conviene considerare il classico modello lineare: $y = X\beta + \varepsilon$, in cui i coefficienti vengono stimati minimizzando la quantità:

$S = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$, da cui deriva la stima $\hat{\beta} = (X^T X)^{-1} X^T y$. Quando la multicollinearità è presente questa stima non ha un'unica soluzione e, di conseguenza, molti o tutti i coefficienti espressi dal vettore β risultano molto elevati. Hoerl e Kennard (1970) risolvono questo problema utilizzando come quantità da minimizzare la somma dei quadrati dei coefficienti di regressione modificati come segue:

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

Il secondo termine prende il nome di penalità. Esso controlla il valore dei coefficienti β, facendo sì che non assumano valori troppo elevati La stima di β risulterà pertanto:

$$\hat{\beta} = (X^T X + I\lambda)^{-1} X^T y$$

in cui *I* indica la matrice identità. Si può dimostrare che le uniche soluzioni possibili sono ottenute per λ>0.

Ora il nostro scopo è quello di applicare il concetto di penalizzazione a risposte binomiali. Si ricorda che:

$$\eta_i = \log \frac{p_i}{1 - p_i} = \alpha + \sum_{j=1}^{p} \beta_j x_{ij}$$

in cui p_i è la probabilità di osservare Y_i=1. Essendo Y_i distribuita come una Bernulliana di parametro p_i si avrà:

$$L(\underline{\beta}) = \prod_{i=1}^{n} P_{Y_i}(y_i, p_i) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

$$l(\underline{\beta}) = \sum_{i=1}^{n}[y_i \log p_i + (1-y_i)\log(1-p_i)] =$$

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

Una la log-verosimiglianza penalizzata può essere definita come:

$$l^*(\underline{\beta}) = l(\underline{\beta}) - \lambda \sum_{j=1}^{n} \beta_j^2 / 2 .$$

Il secondo termine viene detto *ridge penality*; il termine λ regola la penalità: più elevato è il valore di λ, maggiore è la sua influenza e di conseguenza gli elementi di β sono più piccoli. La divisione per 2 è solo una convenienza: questo termine verrà eliminato derivando la log-verosimiglianza.

Da $\frac{\partial l^*}{\partial \alpha} = 0$ *e* $\frac{\partial l^*}{\partial \beta} = 0$ si ottiene un sistema di equazioni di verosimiglianza penalizzata:

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

in cui *u* è il vettore *p*-dimensionale di uno. Le equazioni non sono lineare data la non-linearità tra *p* e α e β. Con l'espansione al primo ordine attraverso la formula di Taylor si ottiene:

$$p_i \approx \tilde{p}_i + \frac{\partial p_i}{\partial \alpha}(\alpha - \tilde{\alpha}) + \sum_{j=1}^{p} \frac{\partial p_i}{\partial \beta}(\beta_j - \tilde{\beta}_j)$$

dove la tilde indica una soluzione approssimata delle equazioni della verosimiglianza penalizzata.

Considerando:

$$\frac{\partial p_i}{\partial \alpha} = p_i(1-p_i)$$

$$\frac{\partial p_i}{\partial \beta} = p_i(1-p_i)x_{ij}$$

e definendo $w_i = p_i(1-p_i)$ e $W = diag(w_i)$, si ottiene:

$$u^T\tilde{W}u\alpha + u^T\tilde{W}X\beta = u^T(y - \tilde{p} - \tilde{W}\tilde{\eta})$$

$$X^T\tilde{W}u\alpha + (X^T\tilde{W}X + \lambda I)\beta = X^T(y - \tilde{p} - \tilde{W}\tilde{\eta})$$

Applicando metodi iterativi si arriva ad una soluzione in maniera generalmente veloce: nella maggior parte dei casi 10 iterazioni sono sufficienti. I valori iniziali di $\tilde{\alpha}$ e $\tilde{\beta}$ sono generalmente scelti come $\tilde{\alpha} = \log\left[\frac{\bar{y}}{1-\bar{y}}\right]$ e $\tilde{\beta} = 0$, con $\bar{y} = \sum_{i=1}^{n} y_i / n$.

Si noti che il coefficiente α è strettamente legato alla frazione di soggetti che presentano la caratteristica di essere appartenenti alla classe 1. Infatti dal differenziale penalizzato calcolato rispetto a questo parametro $\frac{\partial l^*}{\partial \alpha} = 0$ segue che:

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

Il parametro λ di penalizzazione ha un'importanza fondamentale e di solito viene scelto in modo tale che minimizzi il criterio di *Akaike (AIC).* Tale criterio ha la caratteristica di studiare la bontà predittiva di un modello sottraendo il numero di parametri della log-verosimiglianza massimizzata, trovando così un equilibrio tra la complessità del modello e la sua fedeltà ai dati.

Il realtà questo problema del parametro di penalizzazione non è stato ancora ampiamente né accuratamente sviluppato. L'obiettivo che ci si è prefisso è di tentare di approfondire il problema di tale stima utilizzando i criteri conosciuti quali l'AIC (Akaike Information Criterion) e il BIC (Bayesian Information Criterion).

3.3 Stima del parametro di penalizzazione

3.3.1 Il criterio AIC (Akaike Information Criterion)

Il criterio AIC (*Akaike Information Criterion*), è noto nella letteratura statistica come il criterio basato sull'informazione di Akaike. Esso costituisce la base fondamentale tra diverse alternative e modello generatore dei dati.

Il criterio deriva dalla minimizzazione della quantità di informazione, così definita:

$$\hat{I}(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x,\hat{\theta}(y))}\right) dx$$

in cui *f* indica il vero modello che ha generato i dati, *g* è il modello che si vuole adattare ai dati e che si confronta con *f* e $\hat{\theta}(y)$ è la stima di massima verosimiglianza del parametro θ, che indicizza il modello *g.*

Così $g(x, \theta)$ è il modello che approssima i dati , mentre $g(x, \hat{\theta})$ è il modello che approssima i dati quando non si hanno a disposizione i parametri del modello ma delle loro stime (calcolate solitamente con il metodo dei minimi quadrati o con la massima verosimiglianza).

Sviluppando l'espressione da minimizzare otteniamo, per le note proprietà dei logaritmi:

$$\hat{I}(f,g) = \int f(x)\log f(f,x)dx - \int f(x)\log g(x,\hat{\theta})dx$$

e, passando al valore atteso:

$$\hat{I}(f,g) = E_x[\log f(X)] - E_x[\log g(X;\hat{\theta})].$$

Passando al secondo valore atteso otteniamo

$$E_Y[\hat{I}(f,g)] = \int f(y)\left[\int f(x)\log\left(\frac{f(x)}{g(x,\hat{\theta})}\right)dx\right]dy.$$

Dopo alcuni passaggi il risultato che evidenziamo è che il modello che fornisce la massimizzazione della funzione:

$$T = E_Y\left[E_x\left[\log x(X;\hat{\theta}(Y))\right]\right] \tag{3.1}$$

risulta essere il miglior modello.

Akaike (1973) ha mostrato che la verosimiglianza massimizzata è uno stimatore distorto del criterio di selezione. Ha evidenziato inoltre che in caso di modelli annidati tale distorsione è pari al numero dei parametri da stimare (che chiameremo *h*) nel modello considerato come approssimazione di quello vero.

Uno stimatore non distorto del criterio di selezione è:

$$\hat{T} = \log L(\hat{\theta};y) - h = \ell(\hat{\theta}) - h \tag{3.2}$$

o equivalentemente, moltiplicando $\hat{T}$ per la costante -2, si ottiene

$$AIC = -2\log L(\hat{\theta};y) + 2h = -2\ell(\hat{\theta}) + 2h,$$

dove $L(\hat{\theta};y)$ è la verosimiglianza e $\ell(\hat{\theta})$ è la log-verosimiglianza corrispondente.

Il primo addendo che appare, $-2\log L(\hat{\theta};y)$ è indice di bontà di adattamento del modello scelto come approssimazione ad un particolare insieme di dati; mentre il secondo addendo +2*h*, rappresenta una penalizzazione dovuta all'aumentare del numero dei parametri.

Un altro modo di scrivere la formula AIC è il seguente:

$$AIC = Dev(y \mid \hat{p}) + 2m(\hat{\theta}),$$

dove *Dev( )* si intende la devianza che è uguale a $-2\log L(\hat{\theta};y)$ mentre $m(\hat{\theta})$ è l'effettiva dimensione del modello. Nella stima del parametro di penalizzazione la dimensione del modello

non corrisponde alla lunghezza del vettore dei parametri. Hastie e Tibshirani rif [27] propongono di stimarla con:

$m(\hat{\theta}) = traccia(Z(Z^T WZ + \lambda P)^{-1} WZ^T),$

dove Z=[u|X] e P è la matrice identica $p+1 \times p+1$ con l'elemento p_{11} uguale a zero.

3.3.2 Il criterio BIC (Bayesian Information Criterion)

Come già accennato nel paragrafo introduttivo di questo capitolo si considera un criterio molto diffuso per unità sia teoriche che pratiche, *"parente stretto"* del criterio AIC: il criterio BIC (Bayesian Information Criterion).

Esso è definito come

$$BIC = -2\log L(\hat{\theta}; y) + p\log(n)$$

3.4 Regressione multinomiale logistica penalizzata

Data la presenza nel dataset in studio di tre classi, lo scopo è quello di generalizzare la regola di classificazione, ottenuta attraverso la regressione logistica riferita a variabili risposta di tipo dicotomico, e riformularla in riferimento ad una distribuzione, non più binomiale, bensì multinomiale.

Sia dunque:

$y_i = (y_{i1}, y_{i2}, ..., y_{iJ}),$ con

$$y_{ij} = \begin{cases} 1 \quad se \quad y_{ij} = j, \\ 0 \quad se \quad y_{ij} \neq j, \end{cases}$$

e $\sum_j^J y_{ij} = 1$. Sia $p_i = \Pr(Y_i = y_i)$

Dato che Y_i è la realizzazione di una variabile casuale multinomiale, si ponga:

$$p_j(x) = \frac{e^{\underline{\beta}_j^T \underline{x}}}{1 + \sum_{h=1}^{J-1} e^{\underline{\beta}_h^T \underline{x}}}$$

con $\underline{\beta}_j^T = (\alpha_j, \beta_{j1}, ..., \beta_{jp})$ e

$\underline{x} = (1, x_1, ..., x_p)$.

Sia inoltre $p_J = 1 - \sum_{i=1}^{J-1} p_i$

le funzioni di verosimiglianza e log-verosimiglianza, che dopo la penalizzazione risulteranno utili per la stima del vettore dei parametri di regressione.

$$L(\underline{\beta}) = \prod_{j=1}^{J} P_{Y_i}(y_{ij}, p_j(x_i)) = \prod_{j=1}^{J} p_j(x)^{y_{ij}}$$

$$\ell(\underline{\beta}) = \sum_{j=1}^{J-1} y_{ij} \log p_j(x_i) + \left(1 - \sum_{j=1}^{J-1} y_{ij}\right) \log\left[1 - \sum_{j=1}^{J-1} p_j(x_i)\right]$$

$$= \sum_{j=1}^{J-1} y_{ij} \log \frac{p_j(x_i)}{1 - \sum_{j=1}^{J} p_j(x_i)} + \log\left[1 - \sum_{j=1}^{J-1} p_j(x_i)\right]$$

Tramite la funzione di log-verosimiglianza, si può ottenere quella di log-verosimiglianza penalizzata e procedere nella scelta della parametri di regressione e di conseguenza nella scelta della regola di classificazione in maniera del tutto simile al caso di sole due classi descritto nel paragrafo precedente.

3.5 Applicazione ai dati

La regressione multinomiale logistica penalizzata è stata applicata nel dataste descritto nel Capitolo 2.

Per prima cosa si necessita di individuare un valore conveniente del parametro di penalizzazione λ. Per ottenere buoni risultati si è costruita la funzione di log-verosimiglianza penalizzata come descritto nel paragrafo precedente ed, attraverso l'ausilio della funzione *optim* presente in R si sono stimati diversi valori della log-verosimiglianza penalizzata al variare di λ. Per cercare un valore ottimale, tale parametro è stato fatto variare tra 1 e 10^5 usando 25 valori equispaziati del logaritmo in base 10 di λ. Il valore ottimale utilizzando il metodo di Akaike come in Figura 3.1, ottenuta tramite l'ausilio della funzione *curve*, sarà all'incirca pari a 290.

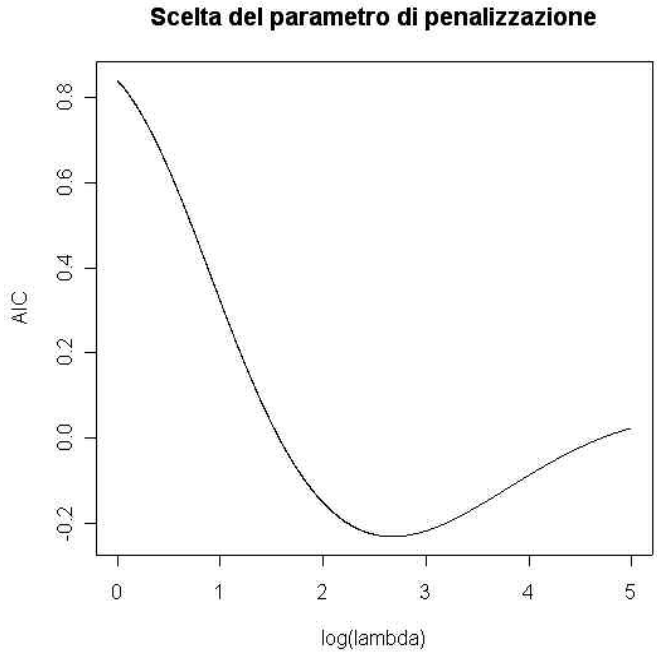


Figura 3.1 Indice AIC al variare del coefficiente di penalizzazione

Si è poi utilizzato il criterio BIC:

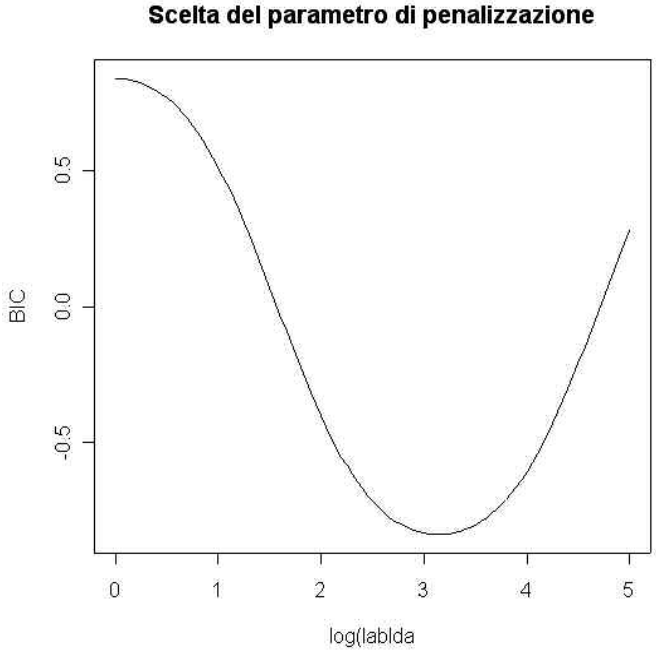


Figura 3.2 Indice BIC al variare del coefficiente di penalizzazione

Si nota che utilizzando lo stesso metodo utilizzato per l'AIC il valore ottimo questa volta è all'incirca pari a 320.

Fissati quindi i due valori di λ ricavati, nella fase successiva si è costruita una regola di classificazione su di una parte del dataset completo (training set), per vagliarlo sulla rimanente (test set). Utilizzando quindi la stima dei parametri ottenuta per i due valori di le seguenti matrici di λ selezionati, si sono ottenute confusione:

Regressione logistica penalizzata con λ ricavato dal criterio Akaike				
Trainin set		Classe prevista		
		ALL	MLL	AML
Classe reale	ALL	20	0	0
	MLL	1	16	0
	AML	0	1	19
Tasso d'errore: 0.035				

Tabella 3.3

Regressione logistica penalizzata con λ ricavato dal criterio Akaike				
Test set		Classe prevista		
		ALL	MLL	AML
Classe reale	ALL	3	0	1
	MLL	0	3	0
	AML	2	1	5
Tasso d'errore: 0.26				

Tabella 3.4

Tabella 3.3 e Tabella 3.4 Matrici di confusione relative all'applicazione della regressione logistica penalizzata con il parametro scelto secondo il criterio Akaike, la parte destra (Tabella 3.3) si riferisce al training set, la parte sinistra (Tabella 3.4) si riferisce al test set.

I risultati sono abbastanza buoni: le probabilità di errore sono molto limitate soprattutto per quanto riguarda il *training set* in quanto si utilizza una numerosità abbastanza superiore al *test set*, ottenendo così una maggior precisione.

Si confrontano poi i risultati ottenuti dalla regola di classificazione a partire dalla penalizzazione trovata con il metodo BIC:

Regressione logistica penalizzata con λ ricavato dal criterio BIC				
Trainin set		Classe prevista		
		ALL	MLL	AML
Classe reale	ALL	19	0	1
	MLL	2	14	1
	AML	0	2	17
Tasso d'errore: 0.10				

Tabella 3.5

Regressione logistica penalizzata con λ ricavato dal criterio BIC				
Test set		Classe prevista		
		ALL	MLL	AML
Classe reale	ALL	2	1	1
	MLL	0	3	0
	AML	1	1	6
Tasso d'errore: 0.26				

Tabella 3.6

Tabella 3.5 e Tabella 3.6 Matrici di confusione relative all'applicazione della regressione logistica penalizzata con il parametro scelto secondo il criterio BIC, la parte destra (Tabella 3.5) si riferisce al training set, la parte sinistra (Tabella 3.6) si riferisce al test set.

In questo caso il tasso di errore aumenta rispetto alla metodologia precedente se si prende in considerazione il training set, rimane invece immutata se si prende in considerazione il test set.

Il tasso di errore non supera mai il 30%. Questo tipo di analisi discriminante sembra quindi abbastanza idonea al tipo di dati considerato.

3.5.1 Un confronto con il metodo *"shrunken centroids"*

La tecnica utilizzata nel capitolo precedente come soluzione del problema dell'elevata dimensionalità, viene qui riproposto come un'ulteriore tecnica di classificazione: si procede quindi con l'applicazione degli algoritmi dell'appendice 2 per avere un termine di paragone con la regressione logistica penalizzata.

Si procede quindi con la scelta del parametro di penalizzazione nel *training set* e con la verifica della classificazione attraverso *cross-validation* nel *test set*:

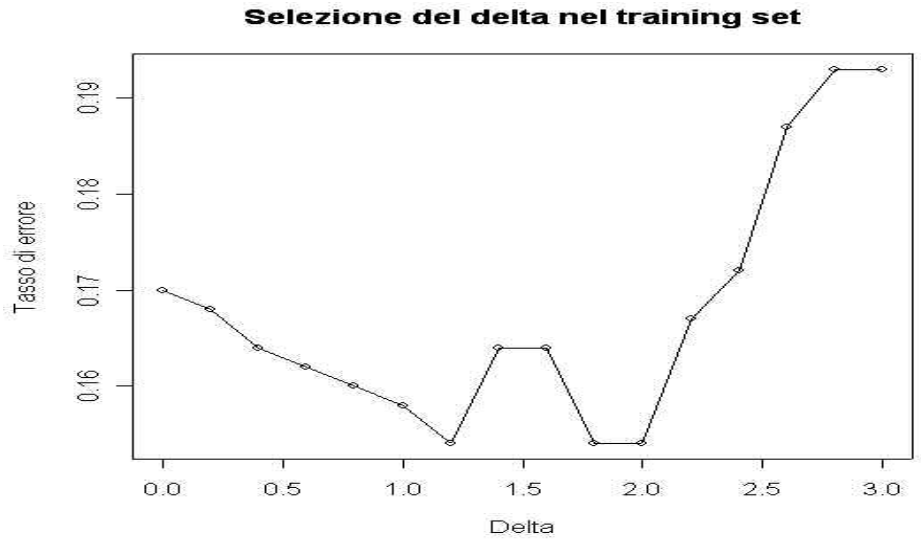


Figura 3.7 Efficacia dell'algoritmo nel training set per diversi valori del parametro Δ

La figura 3.7 mosta la prestazione dell'algoritmo al variare del parametro Δ. Da questa si può notare che il minimo errore si ottiene ponendo Δ=1.2, Δ=1.9 e Δ=2.0. E' utile notare che l'errore prima di Δ=2 decresce, superata questa soglia cresce in maniera abbastanza regolare. E' stato scelto il valore Δ=2 che genera un errore pari a circa 0.15.

Se si applica infatti il metodo *"shrunken centroids"* al training set, la matrice di confusione risulta:

Regressione logistica penalizzata con il metodo shrunken centroids				
Trainin set		Classe prevista		
		ALL	MLL	AML
Classe reale	ALL	17	0	3
	MLL	2	14	1
	AML	1	2	16
Tasso d'errore: 0.157				

Tabella 3.7 Matrici di confusione relativa all'applicazione del metodo *shruken centroid* con la validazione incrociata *leave-one-out* nel *training set.*

Il tasso di errore risulta più elevato rispetto al metodo della regressione logistica penalizzata. Ma per ottenere un'ulteriore prova, si applica l'algoritmo degli *shrunhen centroids* al test set e si confrontano ancora una volta i tassi di errore:

Regressione logistica penalizzata con λ ricavato dal criterio BIC				
Test set		Classe prevista		
		ALL	MLL	AML
Classe reale	ALL	1	2	1
	MLL	1	2	0
	AML	1	0	7
Tasso d'errore: 0.33				

Tabella 3.7 Matrici di confusione relativa all'applicazione del metodo *shruken centroid* con la validazione incrociata *leave-one-out* nel *test set.*

Anche in questo caso si può notare che il metodo della regressione logistica penalizzata ha una probabilità di sbagliare minore rispetto all'algoritmo qui discusso, sebbene sia comunque un ottimo strumento per il controllo della dimensionalità.

APPENDICE CAPITOLO 3

rut.3.1 Funzione che calcola la log-verosimiglianza penalizzata con *"ridge penalization"* di una multinomiale:

```
loglik.pen<-function(beta,data,lambda){
xnam<-paste("x",1:(ncol(data)-1),sep="")
colnames(data)<-c(xnam,"classe")
formula<-as.formula(paste("classe~",paste(xnam,collapse="+")))
m<-model.frame(formula,data)
x <- model.matrix(formula, m)
class.ind<-function(cl) {
       n <- length(cl)
       x <- matrix(0, n, length(levels(cl)))
       x[(1:n) + n * (as.vector(unclass(cl)) - 1)] <- 1
       dimnames(x) <- list(names(cl), levels(cl))
       x
}
attach(data)
classe<-factor(classe)
y<-class.ind(classe)
g<-ncol(y)
pi<-matrix(0,nrow=ncol(y),ncol=nrow(x))

den<-c(rep(1,nrow(x)))
beta<-matrix(beta,nrow=ncol(x),ncol=(g-1))
beta0<-c(rep(0,ncol(x)))
beta<-cbind(beta0,beta)

pi<-matrix(0,nrow=nrow(x),ncol=ncol(y))
for(h in 1:ncol(y)){
pi[,h] <- as.vector(exp( x %*% beta[,h])/(1 + exp( x %*% beta[,2])+exp(  x %*%
beta[,3])))
}

loglik<-0
for(h in 1:ncol(y))
loglik<-loglik+sum(y[,h]*log(pi[,h]))


loglik-lambda*sum(beta^2)
}
```

rut.3.2 Log-verosimiglianza negata da utilizzare nel comando optim presente in R:

```
loglik.pen.neg<-function(beta,data,lambda){
xnam<-paste("x",1:(ncol(data)-1),sep="")
colnames(data)<-c(xnam,"classe")
formula<-as.formula(paste("classe~",paste(xnam,collapse="+")))
m<-model.frame(formula,data)
x <- model.matrix(formula, m)
class.ind<-function(cl) {
       n <- length(cl)
       x <- matrix(0, n, length(levels(cl)))
       x[(1:n) + n * (as.vector(unclass(cl)) - 1)] <- 1
       dimnames(x) <- list(names(cl), levels(cl))
       x
}
attach(data)
classe<-factor(classe)
y<-class.ind(classe)
g<-ncol(y)
```

```
pi<-matrix(0,nrow=ncol(y),ncol=nrow(x))

den<-c(rep(1,nrow(x)))
beta<-matrix(beta,nrow=ncol(x),ncol=(g-1))
beta0<-c(rep(0,ncol(x)))
beta<-cbind(beta0,beta)

pi<-matrix(0,nrow=nrow(x),ncol=ncol(y))
for(h in 1:ncol(y)){
pi[,h] <- as.vector(exp( x %*% beta[,h])/(1 + exp( x %*% beta[,2])+exp(  x %*%
beta[,3])))
}

loglik<-0
for(h in 1:ncol(y))
loglik<- loglik-sum(y[,h]*log(pi[,h]))


loglik+lambda*sum(beta^2)
}
```

rut.3.4 Funzione che calcola l'AIC a partire dai diversi valori che la log-verosimiglianza assume al variare di lambda:

```
aic<-function(ll,data){
aic<-rep(0,length(ll))
for(i in 1:length(ll)){
aic[i]<--2*ll+2*(ncol(data)-1)
}
aic
}
```

rut.3.5 Funzione che calcola il BIC a partire dai diversi valori che la log-verosimiglianza assume al variare di lambda:

```
bic<-function(ll,data){
bic<-rep(0,length(ll))
for(i in 1:length(ll)){
bic[i]<--2*ll+(ncol(data)-1)*log(nrow(data))
}
bic
}
```

Capitolo 4

Considerazioni conclusive

Nel corso del seguente elaborato sono stati analizzati dati riguardanti l'espressione genica di 72 pazienti affetti da 3 diverse forme di leucemia. L'analisi era finalizzata alla valutazione in termini predittivi della tecnica di regressione multinomiale logistica penalizzata: al fine del riconoscimento tra 3 tipi di leucemia sulla base dei suoi livelli di espressione genica.

Dall'analisi condotta emerge che, dal punto di vista statistico, è possibile, grazie all'ausilio di algoritmi costruiti *ad hoc*, realizzare una selezione preliminare delle variabili (ossia geni). Dal punto di vista biologico, si è posto l'accento sull'importanza di individuare ampie classi di geni responsabili della patologia e/o del suo differenziarsi in varie forme. La ricerca e la selezione di geni differenzialmente espressi è stato pertanto uno dei primi obiettivi di questo elaborato.

Si sono condotte analisi di tipo esplorativo: le curve di Andrews sono riuscite ad evidenziare gruppi di geni con diversi livelli di espressione genetica; l'analisi di raggruppamento ha consentito di ricercare gruppi omogenei di geni all'interno di ciascuna patologia, in modo da mettere in luce se geni facenti capo a diverse patologie fossero "simili" in termini di espressione genica. I risultati di questa prima analisi sottolineano la presenza di gruppi abbastanza consistenti di geni con elevato potere discriminante.

Nella fase successiva si è sperimentata la tecnica della regressione logistica penalizzata, in grado di trattare un dataset con le caratteristiche note nella letteratura come *"large p and small n"*. Tale tecnica risulta avere una buona rilevanza dal punto di vista statistico, in quanto non ha un elevato tasso di errore nella classificazione, ma dal punto di vista biologico non ha portato alla luce risultati significativi in quanto non occupa della selezione, ma unicamente della classificazione. Tale tecnica ha risolto però, in un certo senso, il problema dell'elevata dimensionalità che provoca sovraparametrizzazione e multicollinearità nella fase di stima della regola discriminante.

Tale tecnica è stata confrontata con l'algoritmo noto in letteratura com el'algoritmo di Hastie e tibshirani

Complessivamente si può affermare che le tecniche analizzate hanno due aspetti positivi diversi: l'algoritmo proposto da Tibshirani è in grado di selezionare le variabili senza provocare errori di classificazione tanto elevati, quindi ha una grossa rilevanza a livello biologico; la tecnica della regressione multinomiale logistica penalizzata ha la caratteristica di classificare in modo più preciso le osservazioni, quindi ha maggior rilevanza dal punto di vista statistico.

Molte questioni rimangono comunque aperte 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 informazioni significative ma aggiungono soltanto rumore di fondo. Questo tipo di problema fino a poco tempo fa non era molto sentito, ma in questo ambito assume proporzioni consistenti. Infatti una delle condizioni fondamentali richieste per l'applicazione della maggior parte delle tecniche di analisi è quella di disporre di un numero considerevole di osservazioni rispetto alle variabili da analizzare. 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 darà accesso a dataset più adeguati e a misure più attendibili, cosa che garantirà anche la ricerca statistica delle basi più stabili su cui fondare il proprio lavoro.

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 Measured Error for Gene Expression Arrays* (2001)

[3] B.Lausen *Statistical analysis of genetic distance data* (1999)

[4] Golub, T.R. Slomin, D.K.Tamayo, P.Huard et al *Molecular classification of cancer: class discoveri and class prediction by gene expression monitoring* (1999)

[5] Jen-Michel Clavarie Computational methods for identification of differential coordinated gene expression data (1999)

[6] T.R. Golub, D.K. Slonim, Tamayo, C.Huard, M. Gaasembeek, J.P.Mesirov, H.Coller, M.L.Loh, J.R.Downing, M.A. Caligiuri, C.D. Bloomenfield, E.S Lander *Molecular Classification by Gene Expression Monitoring* (Oct 1999)

[7] J.Platt Fast *training of support vector machines using sequential minimal optimization* (1999)

[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* (2000)

[9] A. Brn-Dor, L.Bruhn, N.Friedman, I.Nachman, M.Schummer, Z.Yakhini *Tissue classification with gene expression profiles* (2000)

[10] A.A. Alizadeh, M.B. Elisen, 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 profiles* (2000)

[11] Perou, C.M. Serlie, T. Elsen, M.B. van deRijn, M. Jeffrey, S.S Rees, C.A Pollack, JR Ross, DT Johnsen, H.Akslen et al *Molecular portraits of human breast tumors* Nature (2000)

[12] van't Veer, LJ Dai, H. vande Vijver, MJ, He, YD Hart, AM, Mao, M. Paterse, HL, van derKooy, K. Marton, MJ, Witteveen, AT, et al *Gene expression profiling predicts clinical outcome of breast cancer* Nature(2002)

[13] Amir Ben-Dor, Laurakey Bruhn, Nir Friedman, Iftach Nachaman, Michèl Schummer, Zohar Yakhini *Tissue classification with gene expression profiles, Proceedings of the fourth annual international conference of Computational molecular biology* (apr 2000)

[14] A.Keller, M.Schummer, L-Hood, W. Ruzzo *Bayesian classification of DNA array expression data* Tecnical report, University of Washington (Aug 2000)

[15] Gen Hori, Masato Inoue, Shin-ichi Nishimura and Hiroyuki Nakahara *Blind gene classification based on ICA of microarray data*

[16] K. Zhang, H. Zhao *Assessing reliability of gene clusters from gene expression data* (2000)

[17] M.K. Kerr, G.A. Churchill *Bootrapping cluster analysis: assessing the reliability of conclusions from microarray experiments* (2001)

[18] Allison, devid B., Gadbury, gary L., Heo, Moonseong, Fernandez, José R., Lee, Cheol-Koo, Prolla, Tomas A. and Weimdruch Richard *A mixture model approach for analysis of microarray gene expression data, Computational Statistics and Data Analysis* (2000)

[19] Lorenz Wenisch *Statistical method for microarray data* (2001)

[20] Robert Tibshirani, Guenther Walther, David Botstein, Patrick Brown *Cluster validation by prediction strength*

[21] Fraley, Chris and Raftery, Adrian E. *Model-based clustering, discriminant analysis, and density estimation*. Journal of the American Statistical Association (2002)

[22] Gengxin Chen, Saied A. Jaradat, Nila Banerjee *Evalutation DNA comparison of clustering algorithms in analyzing as cell gene expression data* (2002)

[23] Chiara Romualdi, Stefano Campanaro, Davide Campagna, Barbara Celegato, Nicola Cannata, Stefano Toppo, Giorgio Valle and Gerolamo Lanfranchi *Pattern recognition in gene expressxion profiling using DNA array: a comparative studiy of different statistical methods applied to cancer classification. Human molecular genetics* (2003)

[24] Erich Huang, Skye H Cheng, Holly Dressman, Jennifer Pittman, Mei Hua Tson, Cheng Fang Horng, Andrea Bild, Edwin S Iversen, Ming Liao, Chii MingChen, Mike West, Joseph R. Nevis, Andrei T. Huang *Gene expression predictors of breast cancer outcames* (Mar 2003)

[25] Michel C O'Neil and Li Song *Neural network of lymphoma microarray data: prognosis and diagnosis near-perfect* (Apr 2003)

[26] I. Steinwart On *the influence of kernel on the generalization ability of the support vector machines* Technical report (2001)

[27] Khan J., Wei, JS. Ringer, M. Saal, LH. Ladanyi, M. Westermann, F.Berthold, F. Schwab, M. Antonescu, CR. Peterson, C& Melzer PS Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks Nature Medicine (2001)

[28] Dhanasekaran, SM. Berrette, TR. Chrosh, D. Shah, R. Veranbally, S. Kurachi, K. Pienta, KJ. Rubin, MA & Chinnalyan, AM Delineation of prognostic biomarkers in prostate cancer Nature 2001

[29] Parmeggiani Giovanni, Carret Elisabeth S., Anbazhagan Ramaswamy, and Gabrielson Edward *A statistical framework for expression-based molecular classification cancer* Journal of the Royal Statistical Society (2002)

[30] Tibshirani Robert, Hasti Trevor, Narasimhan Balasubramanian, Elisen Michael, Sherlock Gavin, Brown Pat and Bostein David *Exploratory sceening of genes and cluster from microarray experiment* Statistica Sinica (2001)

[31] Laura Lazzeroni and Art Owen Plaid *model for gene expression data* Statistica Sinica (2002)

[32] Francesca Ruffino, Giorgio Valentini e Marco Muselli *Metodi di Bagging e di selezione delle variabili per l'analisi dei dati di DNA microarray* (2002)

[33] M.K. Kerr *Linear Models for Microarray Data Analysis: Hiden Similiarities and Differences* (2003)

[34] R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu *Diagnosis of multiple cancer types by shrunken centroids of gene expression* (2001)

[35] Pace, Salvan, Ventura *On a asymptotic relation between modified profile likelihood and Akaike information criterion* Working Paper (2004)

[36] B. Laursen *Statistical Analysis of genetic distance data* (1999)

[37] Hastie T., Tibshirani R. and Friedman J. *The Elements of Statistical Learning. Data Mining, Inference and Oredictions*. Sprimger-Verlag, New York (1991)

[38] Eilers P., Boer J., Van Ommen G., Van Houwelinger H., *Classification of Microarray Data with Penalized Logistic Regression*. Proceedings of SPIE 2001

[39] Margaret S. Pepe, Ary M. Longtony, Garnet L. Anderson and Michel Shummer *Selecting differentially expressed genes from microarray experiments* UW Biostatistic working Paper series (Mar 2003)

[40] Hastie T, Tibshirani R.J., Generalized *additive models for medical research* Statistical Methods in Medical Research 1995

[41] C. Tilstone *DNA microarrays: Vital statistics* Nature 2003

[42] Bozdogan H. *Model Selection and Akaike's Information Criterion (AIC): the general theory and analytical wxtensinons* Psycometrica 1987

[43] Lorenz Wenish *Statistical methods for microarray data* (2001)

[44] Antoniadis A. *Personal communication. A short course on Computional and Statistical Aspect of Microarray Analysis* Milan, May 2003

Principali Siti Internet di Interesse

http://www.r-project.org

http://bmr.cribi.unipd.it

http://www.dchip.org

http://www.genmapp.org

http://cmgm.standford.edu

http://www.microarray.altervista.org

http://www.stat.standford.edu/~tibs/SAM

http://www.cribi.unipd.it

http://www.bio.devidson.edu/courses/genomics/chip/chip.html

http://www.bioconductor.org

http://www.microarray.org/sfgf/jsp/home.jsp

http://group.cribi.unipd.it/~chiara