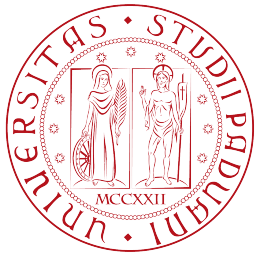

Università degli studi di padova

FACOLTÀ DI INGEGNERIA
Corso di Laurea in Bioingegneria

Integrazione di annotazione funzionale nella classificazione di dati di espressione genica

Relatore:
Di Camillo Barbara

Laureando:
Crepaldi Aler

Correlatore:
Sanavia Tiziana

Anno Accademico
2010/2011

Indice

Capitolo 1

Introduzione: Perchè un nuovo classificatore?

Negli ultimi decenni gli studi genomici hanno fornito un importante supporto alla ricerca medica. In particolare, una delle tecnologie che ha permesso il maggior sviluppo di questo tipo di studi è stata quella dei DNA microarrays che consentono di monitorare, in parallelo, migliaia di molecole e di identificare sequenze geniche caratterizzanti particolari stati fisiologici analizzando i cambiamenti a livello dell'intero genoma.

L'analisi computazionale dei dati provenienti da esperimenti di microarray risulta essenziale per poter formulare nuove ipotesi fisiologiche per rispondere a quesiti sia di tipo diagnostico patologico che funzionale.

In tale contesto i metodi di classificazione supervisionata hanno trovato una vasta applicazione nella genomica moderna. Tali metodologie aprono la prospettiva a diagnosi più realizzabili ed efficienti (Bhattacharjee et al., 2001; Yeoh et al., 2002), alla determinazione del rischio clinico per gruppi diversi di soggetti (Huang et al., 2003; van't Veer et al., 2002) o alla predizione delle risposte a determinati trattamenti (Cheok et al., 2003).

Nell'ambito della classificazione supervisionata applicata ai microarray sono stati proposti diversi algoritmi, dal più classico discriminante lineare (LDA) alle tecniche di penalizzazione (Hui Zou et al., 2005) passando per le support vector machine (SVM) (Vapnik, 1995).

Spesso però i metodi citati prima danno risultati che non hanno diretto riscontro con i meccanismi biologici che si vogliono studiare, tali approcci infatti lavorano su matrici senza considerare l'origine biologica dei dati che si stanno analizzando. I risultati ottenuti quindi sono caratterizzati da liste di geni che risultano spesso poco riproducibili per via delle dimensioni dei dati di microarray. Questo è dovuto essenzialmente al fatto che i metodi di classificazione standard si basano su caratteristiche globali dei dati ma non sulle relazioni biologiche presenti in questi.

Un'altro limite legato ai metodi di classificazione standard è che questi trattano i geni come variabili indipendenti. Informazioni sul ruolo e sulla funzione di molti geni nei relativi processi biologici però sono note in letteratura e opportunamente raccolte in diversi database genomici come ad esempio il database *Gene Ontology* [14].

É possibile riassumere i problemi legati alla classificazioni dei dati di espressione genica appena descritti in due punti chiave:

1. Problemi di dimensione dei dati: I dati forniti da microarray sono caratterizzati da un basso numero di esperimenti (nell'ordine della decina) rispetto ad un alto numero di variabili (migliaia di geni), questo implica la possibilità che il problema di classificazione, che risulta in questo modo fortemente indeterminato, possa avere più di una soluzione;

2. Problemi di correlazione tra i dati: Studi su malattie complesse hanno rivelato come esperimenti su soggetti dello stesso caso clinico siano spesso caratterizzati da alterazioni genetiche eterogenee su più pathway diversi coinvolgendo diverse interazioni tra i geni e l'ambiente cellulare;

Per quanto riguarda il primo punto una soluzione è quella di utilizzare approcci di tipo bootstrap. Per il secondo problema invece sono stati proposti due approcci diversi, un primo approccio consiste nello sviluppo di metodi di classificazioni che tengono conto della correlazione tra le variabili, come ad esempio il metodo pam tra i metodi di classificazione.

Inoltre la ricerca scientifica si sta muovendo nello sviluppo di classificatori in grado di tener conto delle conoscenze sulla biologia che abbiamo a disposizione.

Il presente lavoro di tesi propone una soluzione che si inserisce in questo contesto, il metodo infatti prevede di utilizzare l'informazione biologica contenuta nel database *Gene Ontology* per vincolare la classificazione ad insiemi di geni che appartengono ad uno stesso processo biologico o svolgono la medesima funzione molecolare.

In particolare, l'approccio proposto consente di ridurre il numero di variabili nella definizione del modello di classificazione limitandolo ai soli geni annotati per uno specifico termine dell'ontologia e quindi di ovviare ai problemi legati alla dimensione del dataset.

La tesi è organizzata in quattro capitoli nei quali sono sviluppati i temi principali del lavoro: Integrazione *(capitolo 5)* di annotazione funzionale *(capitolo 4)* nella classificazione *(capitolo 3)* di dati di espressione genica *(capitolo 2)*.

Nel dettaglio, nel secondo capitolo, vengono descritti i dati di espressione genica, illustrando il ruolo dell'rna nella sintesi proteica e quindi nella regolazione della cellula e verrà introdotta la tecnologia dei microarray spiegando come da un campione biologico sia possibile arrivare ad una matrice di dati su cui eseguire analisi.

Nel terzo capitolo invece verrà presentato il problema della classificazione assieme ai metodi maggiormente utilizzati in letteratura nell'analisi di dati da microarray: si partirà in particolare dai metodi per individuare i geni differenzialmente espressi fino ad arrivare a presentare il metodo di classificazione basato su tecniche di regolarizzazione, a cui si farà riferiremo con il nome di *classificatore* l_1l_2, che è stato utilizzato per implementare il nostro algoritmo.

Il quarto capitolo tratta l'informazione biologica e in particolare è focalizzato sulla descrizione del database *Gene Ontology*, uno dei più importanti database in ambito genomico, e dell'utilizzo dell'informazione biologica nella classificazione.

Il quinto capitolo illustra il nostro metodo.

Il cancro come sinonimo di complessità

Il cancro è un complesso di malattie la cui caratteristica fondamentale va ricercata in una crescita cellulare abnorme [5]. Molte sono state le teorie che nei secoli si sono succedute nel tentativo di giungere a questa tanto semplice quanto insidiosa conclusione: dalla teoria degli umori di Ippocrate che si oppone alla concezione antica di malattia come punizione divina, si è passati alle teorie virali di Raus fino a giungere, passando per la teoria cellulare di Wirchow, alla concezione del cancro come malattia che nasce dentro le nostre cellule.

Solamente negli ultimi anni però si è giunti a capire che il cancro è una malattia che deriva da mutazioni somatiche del genoma acquisite da un individuo[1] durante la sua vita.

Questo tipo di mutazione può essere una mutazione a singolo nucleotide, una mutazione su più basi o una mutazione di tipo strutturale. Con l'avvento delle tecnologie di sequenziamento di nuova generazione, come dimostrato da numerosi studi, è possibile avere una misura precisa di queste mutazioni in un gran numero di casi di cancro (Meyerson et al., 2010; International cancer genome consortium, 2010; Mardis and Wilson, 2009).

Alcuni recenti studi [6] sostengono inoltre che il cancro sia la conseguenza del susseguirsi di alterazioni che compromettono sei funzioni primarie della cellula e che la portano ad una proliferazione incontrollata. Secondo Douglas Hanahan e Robert A. Weinberg infatti, una vasta serie di mutazione genetiche che portano al cancro si manifestano sostanzialmente in sei alterazioni funzionali di una cellula. In particolare sostengono che per giungere ad una crescita maligna una cellula deve passare per sei stadi:

1. INDIPENDENZA DAI SEGNALI DI CRESCITA: affinché una cellula normale possa passare da uno stato di quiescenza ad uno stato di proliferazione sono richiesti dei particolari segnali provenienti dall'ambiente esterno, le cellule tumorali auto-generano questi segnali rendendo la crescita indipendente dall'ambiente esterno;

2. INDIPENDENZA DAI SEGNALI INIBITORI DELLA CRESCITA: Il tessuto normale possiede un serie di meccanismi in grado di mantenere le cellule in quiescenza e di garantire l'omeostasi del tessuto stesso, le cellule tumorali evadono questi meccanismi antiproliferazione;

3. PERDITA DI SENSIBILITÀ AI SEGNALI CHE CAUSANO L'APOPTOSI: L'abilità delle cellule tumorali di espandersi non è dovuta solamente all'alto tasso di proliferazione ma anche al fatto che le cellule tumorali sono insensibili ai segnali che causano la morte cellulare;

4. PERDITA DEI LIMITI ALLA REPLICAZIONE: le tre proprietà già acquisite - indipendenza dai segnali di crescita, indipendenza dai segnali inibitori della crescita, perdita di sensibilità ai segnali che causano l'apoptosi - isolano la cellula, dal punto di vista dei processi di crescita cellulare, dall'ambiente circostante. Si è dimostrato però che la distruzione dei meccanismi che regolano lo scambio di segnali cellula-cellula non è sufficiente per

[1] si parla di mutazione somatica, in contrasto a mutazione germinale, quando si verifica una modifica al genoma di una cellula somatica

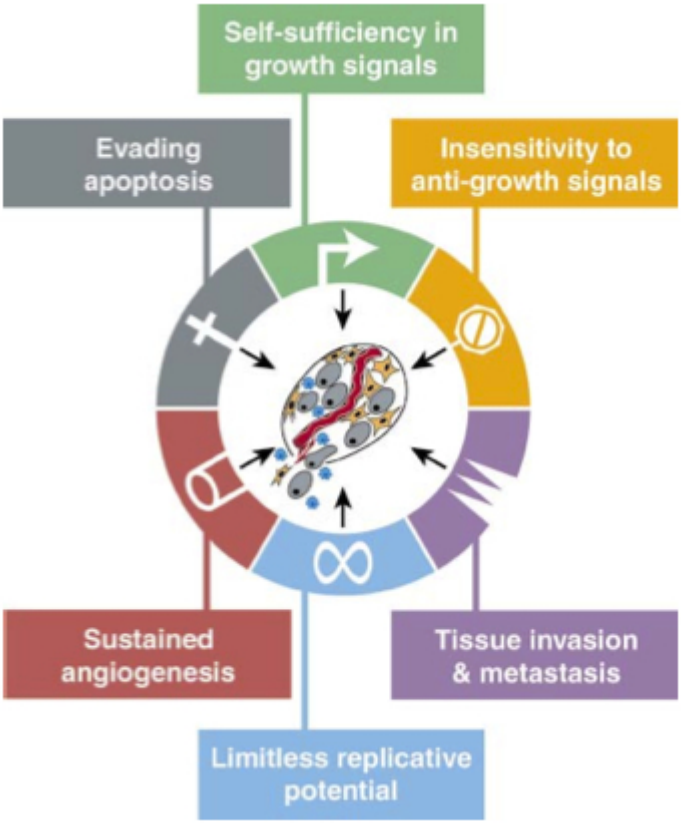

Figura 1.1: Secondo Douglas Hanahan e Robert A. Weinberg il cancro è causato da un evoluzione della cellula in sei fasi, come spiegato nell'articolo le fasi non hanno un ordine preciso e come evidenziato da altri lavori ogni fase potrebbe essere alterata dall'alterazione di uno o più pathway cellulari.

uno sviluppo estensivo del tumore. Molte, o probabilmente tutte le cellule di mammifero possiedono infatti un limite autonomo alla loro replicazione [2]. Affinché il tumore si sviluppi in maniera sensibile anche questo meccanismo deve essere distrutto.

5. PROMOZIONE DELL'ANGIOGENESI: L'ossigeno e i nutrienti portati dalla circolazione sanguigna sono fondamentali per la sopravvivenza e le funzioni cellulari. Per questo, mano a mano che il tumore cresce in dimensioni le cellule tumorali che si allontaneranno sempre di più dai vasi sanguigni saranno costrette a sviluppare meccanismi per l'angiogenesi.

6. SVILUPPO DI MECCANISMI PER LA MOTILITÀ: Prima o poi durante lo sviluppo di molti tipi di cancro la massa tumorale iniziale rilascia delle cellule dotate di motilità, in grado di invadere i tessuti adiacenti e di viaggiare verso parti lontane dal luogo d'origine (metastasi). Queste masse tumorali sparse sono la causa del 90% delle morti per cancro.

Ognuno di questi cambiamenti patologici - che per la cellula hanno il significato di un acquisto di nuove capacità - aprono una breccia ai meccanismi anticancro che le cellule hanno sviluppato in maniera naturale durante la loro evoluzione.

L'ipotesi di Douglas Hanahan e Robert A. Weinberg è avvalorata da altri studi che dimostrano come in realtà un gran numero di mutazioni siano riconducibili all'alterazione di un numero finito di funzioni cellulari.

Ora quindi, come evidenziato anche in [8], la sfida maggiore nell'analisi del genoma di cellule tumorali è quella di distinguere le mutazioni essenziali per lo sviluppo del cancro, *mutazioni primarie* (driver mutations in [8]), da quelle che sono presenti e si accumulano nella cellula ma che non sono causa diretta della malattia, *mutazioni secondarie* (passenger mutation in [8]). Un approccio standard per predire le mutazioni primarie è quello di identificare quei geni che si presentano mutati in un gran numero di casi di cancro.

Questo metodo è stato utilizzato in un lavoro collaborativo mirato allo studio delle mutazioni nel cancro al polmone (vedi [2]). Nell'articolo vengono analizzati 188 casi di cancro e su questi si cercano le mutazioni relative a 623 geni dei quali si conosce la relazione potenziale con il cancro.

L'analisi ha identificato in questo modo 26 geni che risultano mutati in maniera significativa e soprattutto in un gran numero di tumori e che quindi si suppone siano coinvolti nella carcinogenesi.

Questo approccio, pur identificando una gran numero di mutazioni importanti, non è stato però in grado di rivelare tutte le mutazioni principali presenti nei vari pazienti. Studi preliminari inoltre hanno confermato che le mutazioni geniche relative al cancro si presentano come molto eterogenee e non vi sono due genomi - anche se provenienti dallo stesso tipo di tumore - che contengano esattamente le stesse mutazioni somatiche.

Da un articolo di Jones, S et al. emerge ad esempio che il cancro al pancreas non è legato tanto all'alterazione dei singoli geni ma sia conseguenza dell'alterazione di interi pathway. L'articolo, analizzando i profili d'espressione di 24 cellule tumorali, ha messo in luce come vi sia alla base della malattia l'alterazione di 12 pathway significativi prodotta dall'alterazione di geni diversi da caso a caso.

[2] si parla di limite di Hayflick, vedi [7]

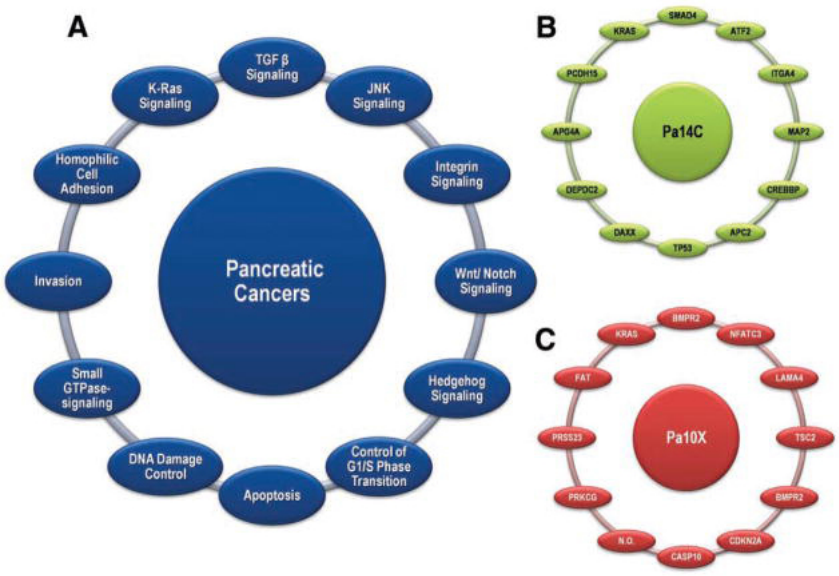

Figura 1.2: A) I 12 pathway e processi significativi per la maggior parte dei tipi di cancro; (B) e (C) Si evidenziano quali geni sono maggiormente corrotti in due particolari tipi di cancro (Pa14C e Pa10X), le posizioni dei geni corrispondono alle posizioni dei pathway.

L'alterazione di questi pathway è condivisa da tutti i 24 tumori analizzati con percentuali che vanno dal 67% al 100% dei casi. Dai risultati dell'articolo si vede ad esempio che il pathway relativo all'apoptosi cellulare risulta compromesso in tutti 24 i tumori analizzati ma ad esempio in un paziente risulta alterato a causa di una mutazione nel gene TP50 mentre in un altro paziente il gene alterato è il CASP10. A livello di singolo gene quindi l'espressione genica risulta diversa ma se si analizzano i risultati a livello di pathways si vede che in realtà le proprietà condivise sono molte di più.

Questo fatto evidenzia come in realtà una stessa malattia genetica possa emergere in pazienti diversi dalla sovra o sotto espressione di geni diversi.

Queste differenze si possono spiegare non solo per la presenza di mutazioni secondarie nei genomi di tutti i pazienti, ma anche grazie all'osservazione che le mutazioni principali colpiscono geni che fanno parte di cascate di segnali o di pathway di regolazione (Hahn and Weinberg, 2002; Vogelstein and Kinzler, 2004). Dato che ognuno di questo pathway contiene molti geni ci sono molte combinazioni di mutazioni principali che possono perturbare un pathway importante per il cancro.

Questa molteplicità di combinazioni rende molto più difficile pensare di poter individuare le mutazioni di tipo funzionale a partire da un calcolo delle frequenze su molti pazienti, pensare di individuare mutazioni principali rare con un metodo del genere richiederebbe un numero di pazienti enorme.

Alla luce delle osservazioni precedenti risulta evidente come il cancro derivi sempre dall'alterazione di una o più cascate di segnali e non dall'alterazione di un singolo gene. Questa affermazione è il motore di tutti gli sviluppi della ricerca nell'ambito della classificazione e dell'analisi di dati di espressione genica.

Molti dei metodi di classificazione proposti dalla letteratura per migliorare l'interpretabilità dei risultati sono basati su di una annotazione *a posteriori* dei geni selezionati in modo da descrivere i processi biologici principali emersi. Recentemente, sono stati proposti però dei metodi che integrano l'informazione biologica direttamente nel processo di apprendimento dei classificatori in modo da ottenere modelli predittivi più verosimili e che tengano conto delle relazioni esistenti tra geni. Il problema del molte soluzioni possibili infatti viene affrontato ponendo dei vincoli sulle relazioni tra i nodi. Questo oltre che a semplificare il problema porta a soluzioni coerenti dal punto di vista biologico. Tra gli altri, Tai e Pan [9] hanno proposto un metodo di classificazione che trattasse i geni appartenenti a differenti gruppi funzionali con diversi termini di penalizzazione con cui sono riusciti ad individuare nel processo di classificazione dei pathway specifici per il cancro. Lottaz e Spang [10] invece hanno proposto un metodo per l'analisi di dati di microarray che genera un grafo di classificazione basato sulla GeneOntology che permette di avere risultati di più facile interpretazione.

Capitolo 2

I dati di espressione genica
I DNA microarray

I dati presi in considerazione in questo studio sono dati di espressione genica. In questo capitolo viene descritto qual'è il legame che intercorre tra questo tipo di dati e la biologia. In particolare verrà introdotto il *dogma centrale della biologia* e verrà messo in luce il ruolo dell'RNA come ponte tra DNA e proteine.

Una volta chiarita la biologia che sta alla base dell'espressione genica si introdurrà la tecnologia dei microarray, descrivendo nel dettaglio i microarray Affymetrix, utilizzati per i dati analizzati in questo lavoro.

La parola *gene* è stata introdotta da Johannsen in 1909 per indicare quell'entità che costituisce il patrimonio ereditario della cellula. Beadle e Tatum poi, con i loro studi sul fungo *Neurospora* hanno mostrato come i geni dirigano la sintesi di enzimi in un rapporto 1:1 [1]. Successivamente Oswald T. Avery ha dimostrato che il supporto fisico che permette l'ereditarietà è l'acido desossiribonucleico (DNA).

Una volta scoperta l'importanza del suolo del DNA nella biologia rimaneva da scoprirne la struttura, ed è proprio da tale scoperta che James Watson e Francis Crick vinsero il premio nobel nel 1962, proponendo il noto modello a doppia elica (vedi figura 2).

Dopo aver introdotto il modello a doppia elica Crick, nel 1958, formulò il *dogma centrale della biologia* ossia il paradigma per il quale l'informazione genetica segue un flusso unidirezionale che va dal DNA alle proteine passando per l'RNA. In particolare il passaggio da DNA a proteine avviene essenzialmente in due fasi principali:

TRASCRIZIONE. In questa fase, che avviene all'interno del nucleo della cellula, il DNA viene tradotto in RNA. Questo processo è mediato da alcuni enzimi detti genericamente RNApolimerasi. Inizialmente l'RNApolimerasi si lega ad alcune sequenze del DNA dette *promotrici* che promuovono la sintesi di nuovo RNA che risulterà complementare al filamento di DNA legato alla RNApolimerasi; infine, quando l'enzima incontra determinate sequenze che indicano la fine di un gene, la trascrizione termina.

TRADUZIONE. Una volta che il DNA è stato trascritto in RNA, questo abbandona il nucleo e può avere inizio la fase di traduzione. In questa fase l'RNA che viene detto a questo punto RNAmessaggero (mRNA), entra in contatto con i *ribosomi* che, con l'aiuto di un'altra molecole detta RNA di trasporto (tRNA) procedono alla traduzione: ad ogni tripletta di basi di mRNA (*codone*) viene associato un amminoacido; la sequenza di più amminoacidi costituisce una proteina.

L'espressione genica quindi avviene nel momento in cui il DNA viene trascritto in RNA. Le cellule eucariotiche possiedono dai 2000 ai 60000 geni codificanti ma in un un dato momento i geni espressi sono solo una piccola percentuale del totale. L'insieme dei geni espressi da una cellula è spesso detto *trascrittoma*.

La comparazione dei profili di espressione genica è stata utilizzata da tempo per rispondere a molte domande relativamente a diversi organismi. Per virus e batteri ad esempio l'analisi dell'espressione genica è servita a capire come l'organismo ospite reagisce ad un attacco virale, negli eucarioti invece gli studi sull'espressione genica hanno ad esempio permesso di individuare i geni alla base di alcuni processi cellulari.

Negli ultimi decenni l'espressione genica è stata studiata usando diverse tecniche come *Northen Blotting* [1] o *polymerase chain reaction* con trascrizione inversa. Ognuna di queste tecniche però permette di analizzare un solo trascritto per volta.

[1] Il Northen Blotting è una tecnica che permette di capire se un gene, la cui sequenza deve essere nota, è espresso in un campione biologico. Attraverso questa tecnica è possibile monitorare uno o pochi RNA per volta. Il processo di valutazione prevede di:

1. Prelevare dai campioni l'RNA pulito e, dopo averlo sottoposto a elettroforesi su gel, trasferirlo su una membrana di nylon (si parla di RNA target);
2. La sequenza di DNA corrispondente al gene che vogliamo monitorare (detta *probe*) è prelevata dal DNA,

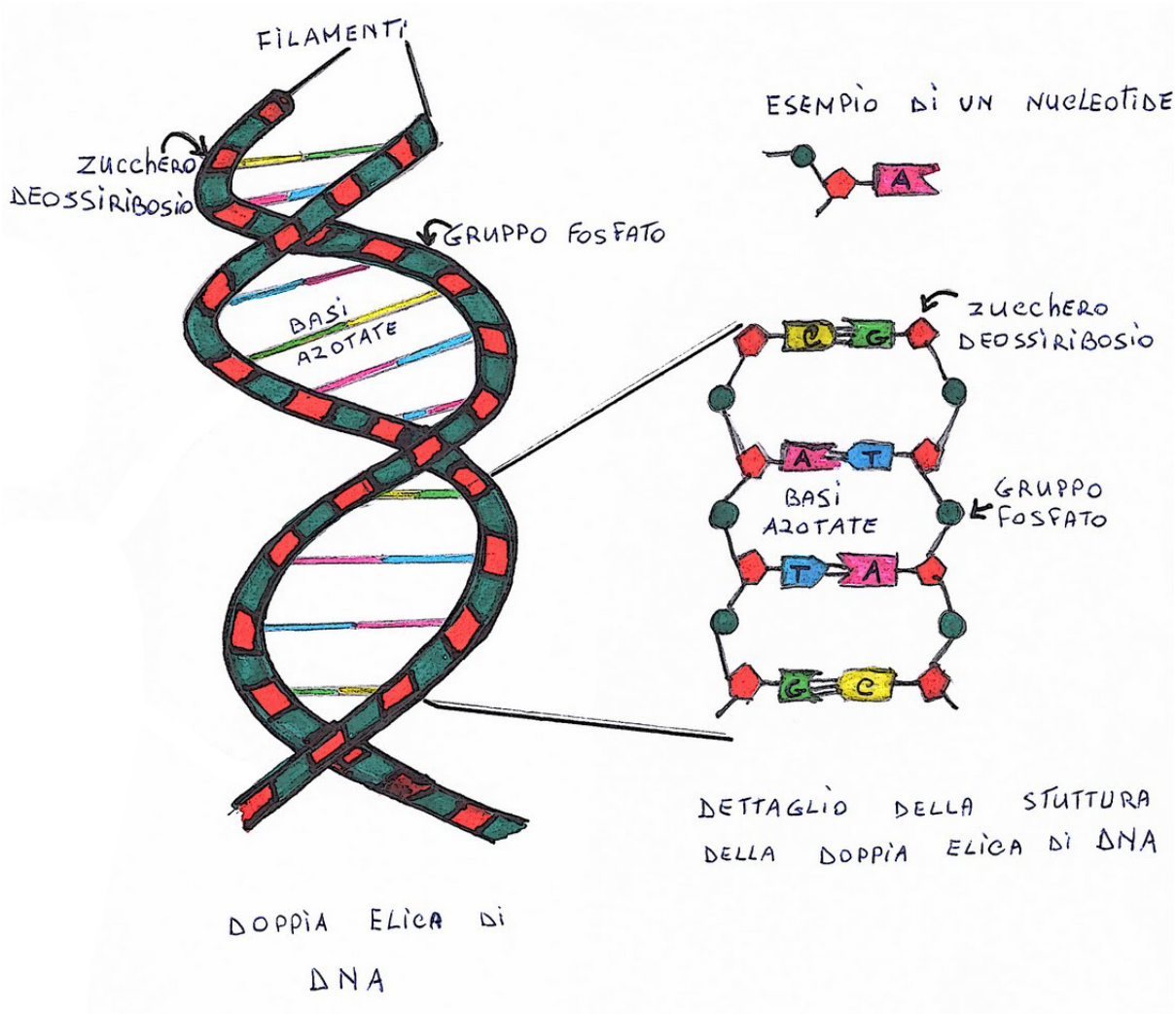

Figura 2.1: Secondo Watson e Crick il DNA è formato da una doppia elica spiralizzata estremamente lunga, questa si presenta simile ad una scala a pioli ruotata a formare una spirale mantenendo i pioli perpendicolari all'asse di rotazione. I due montanti della scala sono formati da molecole alternate di zucchero e fosfato, mentre i pioli sono costituiti da basi azotate legate al tratto di montante ad esse adiacente con un legame covalente. Le basi appaiate sono unite da legami a idrogeno in modo da costituire sempre una coppia purina - pirimidina. Inoltre a causa della struttura delle basi azotate, l'adenina potrà appaiarsi soltanto con la timina mediante due legami a idrogeno (A = T) mentre la guanina soltanto con la citosina, formando 3 legami a idrogeno (G ≡ C). Le basi così appaiate si dicono *complementari*.

In contrasto a tali tecniche si sono successivamente sviluppate altri approcci definiti ad *hight-troughtput*, ossia in grado di monitorare un gran numero di trascritti per volta. Questo approccio offre due importanti vantaggi : 1) può evidenziare la presenza di geni che vengono sovra o sotto espressi in corrispondenza di un particolare stato biologico [2] e 2) può rivelare andamenti di espressione genica legati a particolari stati. Tra le tecnologie hight-troughtput la più diffusa è certamente quella dei *microarrays*.

Un microarray è un supporto solido sul quale vengono depositate in maniera regolare delle sequenze di DNA note (circa 50 basi per ogni sequenza).

L'RNA che viene estratto dal campione biologico di interesse viene spesso retro-trascritto in cDNA mediante l'enzima *trascrittasi inversa*[2] e viene poi marcato con una molecola fluorescente o radioattiva. A questo punto il DNA retro-trascritto viene fatto ibridare al DNA immobilizzato sul microarray. Durante la fase di ibridizzazione il cDNA derivato dalle molecole di RNA del campione biologico di partenza legheranno in maniera selettiva alle sequenze complementari immobilizzate sulla superficie. Successivamente il microarray verrà lavato per eliminare il materiale non ibridato.

Una volta che il microarray è stato ibridato si passa alla lettura del chip. Si eccitano quindi i marcatori i quali emetteranno un segnale la cui intensità verrà rilevata per ciascun trascritto. Un software analizzerà quindi l'immagine risultante e, una volta sottratto il background e calcolata l'intensità che è indicativa del livello di espressione misurato.

I microarray Affymetrics

Il chip Affimetrix consiste il un "wafer" di silicio su cui vengono sintetizzate direttamente sonde composte da sequenze sintetiche di oligonucletidi composti da 20-25 residui.

Il metodo di sintetizzazione in situ consiste nel funzionalizzare il supporto con molecole dette *linker* in grado di agganciare ogni singola base nucleotidica ma con il terminale di reazione bloccato da molecole fotosensibili per una certa lunghezza d'onda. Per rendere disponibili i terminali d'interesse si usa una maschera forata in corrispondenza dei punti in cui deve avvenire la reazione e si illumina il supporto. In questo modo i terminali d'interesse verranno attivati e permetteranno alla base desiderata di agganciarsi. Tale procedimento si ripete per le altre basi con maschere diverse di volta in volta.

Considerando che le dimensione dei fori delle maschere fotolitografiche sono di qualche micron si ha che una sonda corrisponde in realtà a più catene oligonucleotiche tutte uguali (*probe cell*).

---

replicata, marcata e messa in contatto con l'RNA immobilizzato sulla membrana di nylon.

Se il gene è espresso nel campione, il DNA si legherà all' RNA complementare per proprietà di appaiamento delle basi. Il marcatore quindi permette di identificare quali frammenti di DNA si sono legati (ibridati) e all'RNA di quale soggetto.

[2]Il dogma centrale della biologia infatti ammette delle eccezioni rappresentate dai retrovirus. Questi tipi di virus sono dei virus a RNA ma, una volta infettata la cellula ospitante, sono in grado di sintetizzare (*retro-trascrivere*) l'RNA trasportato in DNA che può in alcuni casi fondersi con il DNA della cellula infettata. Il DNA generato a partire dal processo di retrotrascrizione è detto cDNA.

Il limite intrinseco di 20-25 nucleotidi per sequenza, imposto dal processo fotolitografico, rende estremamente difficile ibridare in modo specifico (univoco) una precisa sequenza di RNA. Si può allora ovviare al problema usando più probe complementari per leggere un singolo gene (11, 16 o 20 a seconda del chip usato): in questo modo si garantisce la specificità di ibridizzazione di un particolare blocco di sonde ad un particolare gene. Si parla in questo caso di sonde *perfect match* (PM).

Alle sonde perfect match vengono poi accoppiate delle sonde identiche in sequenza alla sonda corrispettiva di PM tranne che per la base centrale. Infatti spesso accade che le sonde di PM leggano anche sequenze di rumore generate da frammentazioni che non hanno alcuna corrispondenza con il trascritto target, generando così una combinazione di un segnale specifico con uno aspecifico, le sonde errate (con il nucleotide centrale cambiato), dette *mis match* (MM) al contrario leggono solo la parte aspecifica.

Operando la differenza tra i segnali di PM e MM si riesce a ripulire il segnale da un rumore di fondo, inteso come ibridazione aspecifica.

Le 11 sonde di PM sono sparse nel chip, ma sonde PM e MM corrispondenti risultano in celle adiacenti. L'insieme delle sonde PM e MM relative ad un stesso trascritto è detto *probe set*.

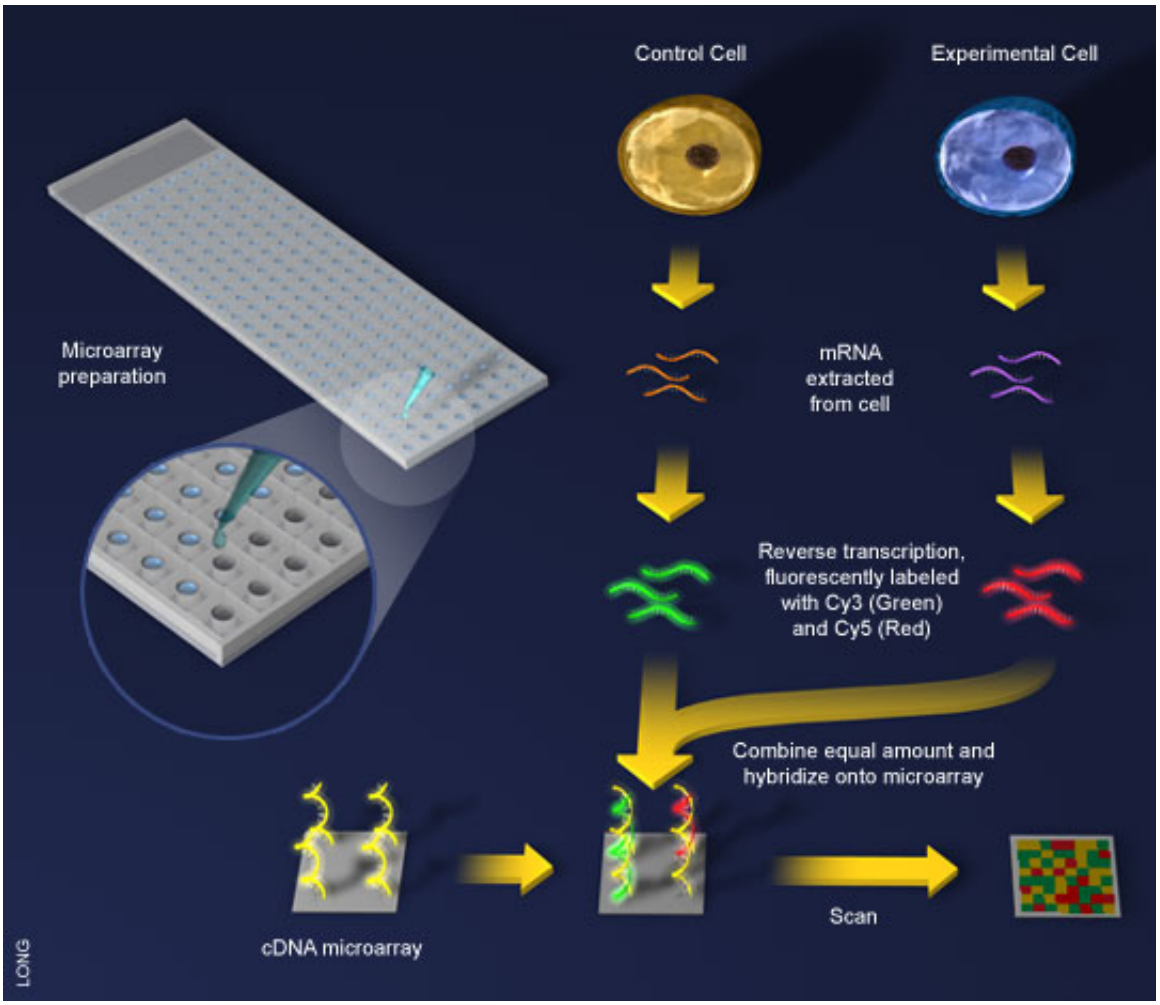

Figura 2.2: Processo di analisi dei dati di espressione genica con microarray Affymetrix.

Capitolo 3

Selezione e classificazione

Nell'ambito dell'analisi di dati di espressione genica sono essenzialmente tre le domande a cui si cerca di rispondere:

1. Quali geni risultano significativamente diversi nei due campioni?
2. É possibile classificare dei soggetti in base alle misure di espressione genica?
3. Quali relazioni tra i geni o tra i soggetti possono essere ricavate dalle misure?

In questo capitolo verranno introdotti dei metodi che permettono di rispondere a queste domande. In particolare nella prima sezione, dedicata ai metodi di selezione, verranno presentati dei test statistici che permettono di individuare, a partire da due classi di soggetti, quali sono i geni che risultano significativamente diversi (*geni differenzialmente espressi*). Dopo aver introdotto il problema quindi vedremo nell'ordine il t-test, il test di Wilkoxon e i test basati sulle permutazioni.

Nella seconda sezione invece si cercherà di rispondere alla seconda domanda presentando dei metodi di classificazione. Dopo una breve introduzione vedremo il metodo PAM, e le *support vector machine*.

Nella terza sezione verrà illustrata una variante del classificatore *elastic-net* che permette di integrare i metodi di selezione ai metodi di classificazione.

Una volta che un classificatore viene progettato si ha la necessità di validarlo, serviranno quindi delle tecniche che, a partire dai dati che abbiamo a disposizione, permettano di calcolare le performance di classificazione. Le ultime due sezioni presentano due di questi metodi: il metodo *bootstrap* e il metodo a *cross-validazione*.

3.1 Metodi di selezione

Una volta che si hanno a disposizione i dati di espressione genica provenienti da un microarray un primo step di analisi si basa sull'identificazione dei geni che sono differenzialmente espressi. Un gene si dice differenzialmente espresso quando, date due classi di soggetti (a.e malati e sani), il livello di espressione nelle due classi è significativamente diverso in senso statistico.

Nel caso in cui una delle due classi sia considerata come classe di riferimento e l'altra come classe test si utilizzeranno espressioni come gene *up-regolato* nel caso in cui il gene sia sovraespresso nella classe test, mentre si parla di gene *down-regolato* nel caso in cui il gene sia sottoespresso.

Diversi strumenti matematici utili per selezionare i geni differenzialmente espressi sono stati proposti, la maggior parte dei quali fa riferimento ai cosiddetti *test di ipotesi*.

Con un test di ipotesi si vuole rispondere ad un quesito relativo alla distribuzione di probabilità di una o più variabili, in una o più popolazioni, a partire dai campioni che si hanno a disposizione. Nel contesto dell'analisi dei dati di espressione genica, esempi di campioni da analizzare sono ad esempio quelli provenienti da acquisizioni effettuate su soggetti affetti da una certa patologia contro quelli relativi ad una popolazione di soggetti sana, oppure campioni relativi a cellule monitorate in uno stato stazionario non perturbato contro campioni di cellule che hanno subito un determinato stimolo nel tempo. Le ipotesi che vengono fatte coinvolgono il confronto di semplici parametri statistici, come ad esempio la media del livello di espressione genica nell'una o nell'altra popolazione.

Indicata con H l'ipotesi che viene fatta su un certo campione si definiscono:

1. H_0: detta *ipotesi nulla* che è quella che prevede che H sia falsa;

2. H_1: detta *ipotesi alternativa* che è quella che prevede che H sia vera;

Il risultato di un test di ipotesi è allora una decisone: *accetto o rifiuto* H_0.

In particolare si ha che ogni qualvolta viene fatta un ipotesi H_0 su un determinato evento[1] si possono verificare i seguenti quattro casi:

1. decido che H_0 è vera (accetto H_0) quando in realtà è falsa

2. decido correttamente che H_0 è vera

3. decido correttamente che H_0 è falsa

4. decido che H_0 è falsa (rifiuto H_0) quando in realtà è vera

nel primo caso si parla di *errore di tipo 1*, i soggetti che ricadono in questa categoria vengono definiti dei *falsi positivi*, nel quarto caso invece si parla di *errore di tipo 2* e si dice che quei soggetti sono dei *falsi negativi*.

Per quanto riguarda l'individuazione di geni differenzialmente espressi in due popolazioni di dati di espressione genica il problema si formalizza nel seguente modo.

I dati possono essere scritti come

$$\left[\begin{array}{cccc||cccc} x_{11} & x_{12} & \dots & x_{1n_a} & y_{11} & y_{12} & \dots & y_{1n_b} \\ x_{21} & x_{22} & \dots & x_{2n_a} & y_{21} & y_{22} & \dots & y_{2n_b} \\ \vdots & & \dots & \vdots & \vdots & & \dots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nn_a} & y_{N1} & y_{N2} & \dots & y_{Nn_b} \end{array}\right]$$

[1] nel nostro caso l'ipotesi è H_0: il paziente é malato

dove x_{ij} è il livello dei espressione genica del gene i nel soggetto j, e con x vengono indicati i soggetti reference mentre con y i soggetti test.

Si vede allora che preso l'insieme di dati relativo ad un gene i si ha che i vettori

$$\left[\begin{array}{cccc} x_{i1} & x_{i2} & \dots & x_{in_a} \end{array} \right] \quad \text{e} \quad \left[\begin{array}{cccc} y_{i1} & y_{i2} & \dots & y_{in_b} \end{array} \right]$$

rappresentano campioni provenienti da due popolazioni. Ad esempio si può assumere che il primo vettore sia un campione di misure di espressione genica prelevate da n_a soggetti proveniente da una popolazione di persone sane, mente si può assumere il secondo vettore come un campione di misure di espressione genica prelevate da n_b soggetti proveniente da una popolazione di persone malate di cancro.

In tale contesto diversi metodi statistici sono stati applicati per la selezione dei geni differenzialmente espressi. In particolare, molti di questi metodi si basano su semplici nozioni di statistica inferenziale (descritti in dettaglio in Appendice 7). Di seguito, si riporta una breve descrizione dei test maggiormente utilizzati.

3.1.1 t-test

Il t-test è un test di ipotesi su due gruppi che si basa sulle seguenti assunzioni:

1. Le osservazioni sui due gruppi sono campioni indipendenti
2. Le popolazioni di partenza hanno distribuzione normale con rispettivamente medie μ_a e μ_b e deviazione standard σ_a e σ_b non note
3. $\sigma_a = \sigma_b = \sigma$

mentre le ipotesi che vengono effettutate sono

$$H_0 : \mu_a = \mu_b \quad , \quad H_1 : \mu_a \neq \mu_b$$

Ricordando allora che (in riferimento ai risultati in Appendice 7)

a. La differenza di due variabili aleatorie normali è ancora una variabile aleatoria normale che ha come media la differenza delle medie e come varianza la somma delle varianze.

b. Data una variabile aleatoria normale sommando a questa la sua media e dividendo per la sua deviazione standard ottengo una variabile aleatoria normale standard[2].

c. Il rapporto tra una variabile aleatoria normale standardizzata e una variabile aleatoria chi-quadro di ordine r è una variabile aleatoria con distribuzione t di Student con un numero di gradi di libertà pari a r.

Definendo con X_{n_a} e Y_{n_b} l'insieme delle n_a, n_b osservazioni delle classi a e b rispettivamente, e con m_a e m_b le corrispondenti medie campionarie risulta[3]

[2]Si variabile normale standard una variabile normale con $\mu = 0$ e $\sigma^2 = 1$.
[3]vedi Appendice 7

$$m_a \sim N\left(\mu_a, \frac{\sigma_a^2}{n_a}\right) \quad , \quad m_b \sim N\left(\mu_b, \frac{\sigma_b^2}{n_b}\right)$$

da cui segue per il punto a

$$m_a - m_b \sim N\left(\mu_a - \mu_b, \frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}\right) = N\left(\mu_a - \mu_b,\ \sigma^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)\right)$$

quindi per il punto b

$$\frac{(m_a - m_b) - (\mu_a - \mu_b)}{\sqrt{\sigma^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}} \sim N(0,1) \tag{3.1}$$

si noti che a questo punto, se le varianze delle due popolazioni fossero note, è possibile applicare un test di ipotesi: infatti supponendo valida l'ipotesi H_0 la 3.1 diventerebbe

$$\frac{(m_a - m_b)}{\sqrt{\sigma^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}} \sim N(0,1) \quad \Rightarrow \quad P\left[\frac{(m_a - m_b)}{\sqrt{\sigma^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}} | H_0\right] = N(0,1)$$

e si avrebbe quindi una stima della probabilità di sbagliare accettando H_0 in funzione della sola differenza delle due medie campionarie e del numero di campioni.

Tuttavia, se non è possibile conoscere a priori le varianze delle due popolazioni, si utilizza un suo stimatore che si può definire con la seguente formula:

$$s_p^2 = \frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{(n_a + n_b - 2)}$$

Si può dimostrare (per la validità del punto c) che la variabile aleatoria y (calcolata considerando valida H_0) risulta

$$y = \frac{(m_a - m_b)}{\sqrt{s_p^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}} \sim t(n_a + n_b - 2)$$

si trova quindi la funzione (dipendente solo da n_a e n_b)

$$P[y|H_0] \sim t(n_a + n_b - 2) \tag{3.2}$$

Tale funzione in particolare mi fornisce la probabilità che ho di trovare un valore y considerando vera l'ipotesi H_0.

Si noti quindi che, considerando di aver calcolato il valore di $y = y^{obs}$ per una particolare coppia di dati, $P[y^{obs}|H_0]$ indica la probabilità che la variabile y sia uguale a y^{obs} sotto l'ipotesi che H_0 sia vera (e che quindi le medie siano uguali). É quindi possibile definire la seguente regola di decisione:

$$\begin{aligned} \text{se } P[y^{obs}|H_0] \geq 0.95 &\rightarrow \text{accetto } H_0 \\ \text{se } P[y^{obs}|H_0] < 0.95 &\rightarrow \text{rifiuto } H_0 \end{aligned}$$

Di conseguenza la probabilità di commettere un errore di tipo 1,ovvero la probabilità di rifiutare H_0 quando H_0 è vera vale $1 - 0.95 = 0.05$. Fissata allora la probabilità α di commettere un errore di tipo 1, è possibile modificare la soglia di decisione come

$$\begin{aligned} \text{se } P[y^{obs}|H_0] \geq 1-\alpha &\rightarrow \text{accetto } H_0 \\ \text{se } P[y^{obs}|H_0] < 1-\alpha &\rightarrow \text{rifiuto } H_0 \end{aligned}$$

Nel caso del t-test, definendo un valore soglia θ tale che

$$P[\ |\theta|\ \ |H_0] = 1 - \alpha$$

le regole definite prima diventano

$$\begin{aligned} \text{se } \left|y^{obs}\right| < \theta &\rightarrow \text{accetto } H_0 \\ \text{se } \left|y^{obs}\right| \geq \theta &\rightarrow \text{rifiuto } H_0 \end{aligned}$$

infatti risulta che

$$\left|y^{obs}\right| < \theta \Rightarrow P[y^{obs}|H_0] \geq P[\theta|H_0] = 1 - \alpha$$

si noti inoltre che per calcolare θ è sufficiente conoscere il valore della funzione t di Student di grado $n_a + n_b - 2$ solamente in corrispondenza di alcuni valori significativi del valore α[4]. Al contrario, per calcolare $P[y^{obs}|H_0]$ è necessario conoscere il valore della t di Student corrispondente a tutti i possibili valori di y^{obs}.

Riassumendo quindi una volta che abbiamo a disposizione i campioni presi dalle due popolazioni e assunte valide le ipotesi descritte in 3.1.1 si procede nel modo seguente:

1. definisco l'ipotesi $H_0\ :\ \mu_a = \mu_b$
2. fisso un livello di significatività α (ad esempio $\alpha = 0.05$)
3. calcolo $\theta\ :\ P[\ |\theta|\ \ |H_0] = 1 - \alpha$ per una variabile t di Student con $n_a + n_b - 2$ gradi di libertà
4. calcolo y^{obs} come

$$y^{obs} = \frac{(m_a - m_b)}{\sqrt{s_p^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}} \quad \text{con} \quad s_p^2 = \frac{(n_a-1)s_a^2 + (n_b-1)s_b^2}{(n_a + n_b - 2)}$$

[4]tipicamente vengono utilizzati $\alpha = 0.01$ e $\alpha = 0.05$

5.

$$\text{se } \left|y^{obs}\right| < \theta \quad \rightarrow \quad \text{accetto } H_0 \text{ (i campioni provengono dalla stessa popolazione)}$$
$$\text{se } \left|y^{obs}\right| \geq \theta \quad \rightarrow \quad \text{rifiuto } H_0 \text{ (i campioni provengono da popolazioni diverse)}$$

Se desideriamo quindi capire quali geni risultano differenzialmente espressi nelle due popolazioni possiamo supporre valide le ipotesi del t-test e cercare quei geni per cui $\left|y^{obs}\right| \geq \theta$. Affinchè un gene possa dirsi differenzialmente espresso le due popolazioni devono poter essere considerate come significativamente diverse e quindi il test viene usato per determinare quando l'ipotesi nulla H_0 viene rifiutata.

3.1.2 test SAM per i microarray

Negli esperimenti con i microarray n_a e n_b sono bassi e quindi s_p non è una stima robusta di σ e questo porta a commettere degli errori, in particolare

1. se θ è sottostimata allora $|y^{obs}|$ risulterà alta
2. se θ è sovrastimata allora $|y^{obs}|$ risulterà bassa

quindi nel caso 1 sarò portato a selezionare più geni come differenzialmente espressi di quanto non farei con una stima corretta di θ, viceversa nel caso 2 ne selezionerei meno. Quello che però mi interessa è di selezionare solo i geni che realmente possono essere considerati differenzialmente espressi, definisco quindi una nuova statistica che possa tenere conto di una eventuale sottostima di θ, definisco

$$t_{SAM} = \frac{m_a - m_b}{\sqrt{s_p^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)} + c}$$

L'idea allora è quella di sfruttare il numero elevato di variabili a disposizione (geni) per stimare il parametro c.

Divido per questo il range di s_p in L intervalli equidistanti, è allora possibile formulare le seguenti ipotesi

1. In ogni intervallo ci sono più geni che non sono differenzialmente espressi che geni differenzialmente espressi
2. t_{SAM} dei geni che non sono differenzialmente espressi devono avere valori simili nei vari intervalli.

Quindi

1. Considero $c = (c_1, \ldots, c_2, c_i, \ldots, c_q)$ che vanno dal massimo al minimo valore di $s_p^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)$ osservati;
2. Calcolo t_{SAM} per ciascun gene

3. Calcolo $MAD(t_{SAM})$ in ognuno degli L intervalli in cui ho diviso il range s_p

4. Calcolo $w_q = CV(v_1, \ldots, v_L)$

Scelgo quindi c^* tale che $c^* = \arg\min(w_q)$.

3.1.3 Wilcoxon rank sum test

Il t-test formula delle ipotesi su due popolazioni utilizzando dei parametri definiti sulle loro distribuzioni, ipotizzate gaussiane, ovvero in base alla media e alla SD. Il test esaminato in questo paragrafo invece serve a verificare ipotesi concernenti il modello stesso che caratterizza una o più popolazioni. Si parla in questo caso di test non parametrici. Un esempio è il *Wilcoxon rank sum test* Le assunzioni su cui si basa il test di Wilcoxon sono

1. Le osservazioni sui due gruppi sono campioni indipendenti

2. Le due popolazioni hanno lo stesso tipo di distribuzione (incognita, non gaussiana)

3. La varianza delle popolazioni è uguale

Dati quindi due campioni $(X_1, X_2, \ldots, X_{n_a})$ e $(Y_1, Y_2, \ldots, Y_{n_b})$ le cui distribuzioni di probabilità siano rispettivamente F_1 e F_2 si ordinano le osservazioni in un'unica lista in ordine crescente e si sostituisce poi ad ogni valore il posto occupato nella graduatoria, cioè 1 al valore più piccolo, 2 al successivo e così via. Questi nuovi numeri si chiamano *ranghi*. Sia W_X la somma dei posti in graduatoria corrispondenti alle osservazioni del primo campione. Il test di Wilcoxon assume come statistica test la quantità W_X.

In termini generali, i numeri naturali da 1 a $n_a + n_b$ che contraddistinguono i posti in graduatoria delle osservazioni dei due campioni $(X_1, X_2, \ldots, X_{n_a})$ e $(Y_1, Y_2, \ldots, Y_{n_b})$ considerati congiuntamente sono contrassegnati dai simboli X e Y a seconda che si riferiscano a valori del primo o del secondo campione. Si ottiene perciò una successione del genere

$$\begin{array}{cccccccccc} 1 & 2 & 3 & \ldots & i & \ldots & n_a+n_b-2 & n_a+n_b-1 & n_a+n_b \\ X & Y & Y & \ldots & X & \ldots & X & Y & X \end{array}$$

dove X è presente n_a volte e Y è presente n_b volte. Se l'ipotesi nulla, $H_0 : F_1(z) = F_2(z)\ \forall z$ è vera, i campi possono considerarsi come provenienti dalla stessa popolazione; perciò, il posto in graduatoria di una osservazione non dipende dal campione a cui essa appartiene. Indicando con $R_1 < R_2 < \cdots < R_{n1}$ i posti in graduatoria delle osservazioni del primo campione, se H_0 è vera, tutte le possibili determinazioni di $(R_1, R_2, \ldots, R_{n_a})$ sono equiprobabili e la probabilità di ciascuna determinazione è $\frac{1}{\binom{n_a+n_b}{n_a}}$, di conseguenza la probabilità che W_X assuma un certo valore c è data dal rapporto

$$P(W_X = c \mid H_0) = \frac{\#(W_x = c)}{\binom{n_a+n_b}{n_a}}$$

Per dimensioni campionarie piccole, le soglie critiche di W_X sono raccolte in tabelle precalcolate. Inoltre, si può dimostrare che la distribuzione limite di W_X è

$$y = \frac{W_X - \frac{n_a(n_a+n_b+1)}{2}}{\sqrt{\frac{n_a n_b(n_a+n_b+1)}{12}}} \sim N(0,1)$$

Riassumendo i passi da fare si ha

1. Calcolo W_X^{obs} come somma dei valori dei ranghi nella classe a
2. Fisso un valore di significatività α
3. Calcolo $p^{obs} \ : \ P[\ |w| \ > W_X^{obs}] = p^{obs}$
4. Se $p^{obs} > \alpha$ allora rifiuto H_0, altrimenti accetto

3.1.4 test basati sulle permutazioni

Nei test basati sulle permutazioni non conosco la distribuzione della statistica test y ma la ricavo permutando i dati.

Voglio testare l'ipotesi nulla H_0 : le due popolazioni hanno la stessa distribuzione.

Il processo, una volta scelta una statistica campionaria y, è il seguente

```
FOR (b in 1:B) :
  1. Permuto i dati sulle colonne
  2. Calcolo y* sui dati permutati
END
```

Con la prima operazione vengno ri-assegnati in maniera casuale i soggetti ai due gruppi, mentre con la seconda operazione calcolo il valore della statistica y sui nuovi campioni. Alla fine del processo ho B valori osservati y^* che mi danno una stima della distribuzione di y quando vale l'ipotesi nulla, riassegnare in maniera casuale i soggetti nelle due classi infatti corrisponde a considerarli come appartenenti ad una stessa popolazione (una popolazione mista).

Una volta avuta una stima della distribuzione di y e fissato un livello di significatività α calcolo il valore osservato di y^{obs} sui dati non permutati e il valore

$$p = \frac{\#\left[|y^*| > |y^{obs}|\right]}{B}$$

quindi se $p < \alpha$ rifiuto H_0 altrimenti la accetto.

3.2 Metodi di classificazione

Se nell'ambito della selezione vengono identificati geni che risultano significativamente diversi nei due campioni nel problema di classificazione si vuole definire un modello in grado di assegnare correttamente campioni ad una particolare classe della popolazione di appartenenza.

Si assume quindi che le variabili da predire possano essere assegnate ad uno dei K valori predefiniti $\{c_1, c_2, \ldots, c_K\}$, i K valori corrispondono a K classi predefiniti e se $K = 2$ si parla di classificazione binaria.

Per ogni soggetto possiamo definire una variabile dipendente y_i, detta *label*, $y_i \in \{1, 2, \ldots, K\}$ e un insieme di G misure indicato con $\mathbf{x} = (g_1, g_2, \ldots, g_G)$ detto *vettore delle feature*. Il vettore delle feature si considera appartenente allo *spazio delle feature* X. I soggetti sono caratterizzati da un'etichetta che li associa alle classi note a priori. Sia quindi $y = y_1, y_2, \ldots, y_M$ la variabile contenente i valori delle label degli M soggetti e sia $X = (x_1, x_2, \ldots, x_G)$ il vettore contente l'insieme dei G geni considerati.

Un classificatore C allora è una funzione tale che

$$\begin{array}{rccl} C: & X & \to & \{1, 2, \ldots, K\} \\ & C(\mathbf{x}) & \mapsto & y \end{array}$$

Si ha che la funzione C in sostanza corrisponde ad una partizione dello spazio delle feature X in K sottoinsiemi disgiunti $A_1, A_2, \ldots, A_K$ per cui un soggetto il cui vettore delle feature $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in A_k$ venga associato alla classe $\hat{y} = k$.

3.2.1 Linear Discriminant Analysis e Predictive Analysis of Microarray

Un classico metodo di classificazione, molto noto per la sua semplicità e rubustezza è il metodo a discriminante lineare (LDA).

Supponiamo quindi di avere una variabile $y_i \in \{1, 2, \ldots, K\}$ e il relativo vettore delle feature $\mathbf{x}_i$, si ha allora che la regola per la decisione ottima per la regola di Bayes risulta:

$$\hat{y}(\mathbf{x}) = \arg\max_{k \in G} P(\mathbf{x}|y = k)P(y = k) = \arg\max_{k \in G} P(y|\mathbf{x})$$

quindi supponendo che $P(\mathbf{x}|y = k)$ abbia una distribuzione normale multivariata, risulta:

$$\hat{y}(\mathbf{x}) = \arg\max_{k} \left[\mathbf{x}^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log\left(\frac{n_k}{n}\right) \right]$$

dove la funzione di costo è una funzione lineare su $\mathbf{x}$. Il problema della classificazione quindi si riduce alla stima dei parametri della funzione di probabilità $P(\mathbf{x}|y = k)$. Tradizionalmente si usano per questo degli stimatori a massima verosimiglianza per μ_k e Σ, si ha

$$\hat{\mu}_k = (\hat{\mu}_{1k}, \hat{\mu}_{2k}, \ldots, \hat{\mu}_{pk})^T \quad \text{con} \quad \hat{\mu}_i^k = \frac{1}{n_k} \sum_{y_j = k} \mathbf{x}_{ij}$$

e

$$\hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{y_j=k} (\mathbf{x}_j - \hat{\mu}_k)(\mathbf{x}_j - \hat{\mu}_k)^T$$

Tuttavia per dati numerosi e con pochi soggetti, come è il caso dei dati da microarray, però il classificatore a discriminante lineare soffre della singolarità della matrice di covarianza e la perdita della possibilità di fare feature selection. Per rimediare a questi problemi, Tibshirani *et al.* nel 2003 proposero una modifica al LDA, proponendo il metodo *nearest shrunken centroid*, noto anche con il nome di *Predictive Analysis of Microarray* (PAM) che assume l'indipendenza tra le variabili in modo da tenere in considerazione il problema della singolarità e utilizza un stimatore a soglia per eseguire la selezione delle feature.

Sia quindi x_{ij} il valore di espressione genica del geni i per il soggetto j. Si assuma poi per semplicità di trattazione che $\bar{x}_i = \sum_j \frac{x_{ij}}{n} = 0$ per ogni gene.

L'idea di base della PAM e di ridurre i baricentri delle distribuzioni delle feature delle varie classi (*centroidi*) $\hat{x}_i$ rispetto al centroide complessivo $\bar{x}_i$. Si definisce quindi la distanza tra due classi come:

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

dove s_i è la deviazione standard intra-classe

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

e s_0 è una costante positiva, usualmente scelta come $\{s_i : i = 1, \ldots, p\}$ e $m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}}$. Possiamo allora riscrivere

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

Si può allora mettere una soglia a zero a d_{ik} con l'aggiunta di una costante $\lambda \geq 0$ arrivando ad una *soft thresholding*:

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

con λ deciso solitamente via cross validazione (3.5). Si ottengono quindi i nuovi centroidi per le classi

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

e si definisce la funzione discriminante come

$$\delta_k(\mathbf{x}) = \sum_{i=1}^{p} \frac{(\mathbf{x}_i - \bar{x}'_{ik})^2}{s_i^2} - 2\log\left(\frac{n_k}{n}\right)$$

con $\mathbf{x}$ test set.

3.2.2 Support vector machine

I classificatori che si basano sulle *support vector machine* (SVM) sono progettati con l'idea di trovare il miglior iperpiano in grado di separare le due classi nel training set.

Il caso più semplice per quanto riguarda le SVM è quello in cui lo spazio delle feature, definito a partire dai vettori delle feature $\mathbf{x}$ dei soggetti appartenenti al trainig set, è *linearmente separabile* ossia quando esiste un iperpiano in grado di partizionare lo spazio in due.

L'iperpiano è definito in modo tale che ogni punto $\mathbf{x}$ dello spazio delle feature che lo costituisce soddisfi la seguente equazione

$$\mathbf{wx} + b = 0$$

Definiamo allora il *margine* di un iperpiano di separazione come la somma delle distanze dall'iperpiano dalla più vicina osservazione nella classe positiva e dalla più vicina osservazione nella classe negativa. L'iperpiano cercato allora sarà, tra tutti i possibili iperpiani separatori, quello con il margine più grande.

Considerato che le classi sono per ipotesi linearmente separabili si ha che è sempre possibile trovare una coppia di iperpiani H_1 e H_2 tali che per $i = 1, \ldots, n$ valga

$$\begin{aligned} H_1 : &\quad \mathbf{wx} + b \geq 1 \quad \text{per} \quad y_i = 1 \\ H_2 : &\quad \mathbf{wx} + b \leq -1 \quad \text{per} \quad y_i = -1 \end{aligned}$$

Si noti per inciso che i due iperpiani sono paralleli e che nessun soggetto del training set cade tra di essi. Le due diseguaglianze precedenti possono essere poi condensate in un unica disequazione

$$y_i(\mathbf{wx} + b) - 1 \geq 0 \quad , \quad \forall i$$

la distanza dei due iperpiani dall'origine poi è data rispettivamente da

$$\frac{|1 - b|}{||\mathbf{w}||} \quad \text{e} \quad \frac{|-1 - b|}{||\mathbf{w}||}$$

quindi il margine, definito come distanza tra gli iperpiani, è dato da

$$\frac{2}{||\mathbf{w}||}$$

e quindi massimizzare il margine equivale a minimizzare $||\mathbf{w}||$. Il problema quindi diventa quello di minimizzare $||\mathbf{w}||^2$ sotto la condizione

$$y_i(\mathbf{wx} + b) - 1 \geq 0 \quad , \quad \forall i$$

Si noti per inciso che i soli punti che condizionano la soluzione del problema sono i soli due punti (uno per classe) per cui vale l'equazione $y_i(\mathbf{wx}+b)-1 = 0$, tali punti sono detti appunto *support vector*.

Applicando allora il metodo dei moltiplicatori di Lagrange il problema può essere riformulato nel modo seguente

$$(\mathbf{w}, b) = \arg\min [L(\mathbf{w}, b, a)]$$

con

$$L(\mathbf{w}, b, a) = \frac{1}{2} ||\mathbf{w}||^2 - \sum_{i=1}^{M} a_i \left[y_i(\mathbf{w}\mathbf{x} + b) - 1 \right]$$

A questo punto è possibile determinare i parametri $\mathbf{w}$ e b che determina l'iperpiano di classificazione, la classificazione risulta quindi

$$y_i = sgn\{\mathbf{w}x_i + b\}$$

La soluzione precedente è stata trovata sotto l'ipotesi di classi linearmente separabili. Se questo non si dovesse verificare si potrebbe cercare una soluzione cosiddetta a *soft margin*.

Si cerca in questo caso una soluzione di compromesso tra la larghezza del margine e il numero di soggetti non classificati correttamente, si ammette cioè che alcuni soggetti del training set non vengano classificati correttamente.

Si introducono allora delle nuove variabili ξ_i dette *variabili di slack* e si va a minimizzare la funzione

$$(\mathbf{w}, b) = \arg\min \left[\frac{1}{2} ||\mathbf{w}||^2 + c \sum_{i=1}^{M} \xi_i \right]$$

sotto la condizione

$$y_i(\mathbf{w}x_i + b) \geq 1 - \xi_i \quad \forall i \quad , \quad \xi_i \geq 0$$

Si vede poi che le variabili di slack hanno il seguente significato

$\xi_i = 0$: il soggetto i è classificato correttamente o giace sulla linea di margine

$0 < \xi_i \leq 1$: il soggetto i giace dentro il margine ma dal lato corretto

$1 < \xi_i \leq 2$: il soggetto i giace dentro il margine ma dal lato errato

$\xi_i > 2$: il soggetto i giace oltre il margine dell'altra classe

quindi si vede che i due termini della funzione costo precedente servono rispettivamente a massimizzare il margine e a minimizzare il numero di soggetti classificati nel gruppo errato.

3.3 Integrazione tra i metodi di classificazione e di selezione: Il classificatore l_1l_2

L'algoritmo di classificazione l_1l_2 è stato introdotto sia per la classificazione di nuovi soggetti sia per la selezione di *variabili rilevanti per eseguire una buona classificazione* [3]. Tale problema risulta particolarmente evidente nell'ambito dell'analisi dei dati di espressione genica caratterizzati dall'elevato numero di geni che vengono monitorati. In questo caso il modello di classificazione quindi sarà descritto da una funzione di predizione C, definita nell'introduzione, che dipende dalla combinazione lineare di alcune variabili solamente.

Nell'analisi di dati di espressione genica un'altro problema da tenere in considerazione è dato dall'altro livello di correlazione che intercorre tra i geni monitorati, che complica ulteriormente l'identificazione robusta dell'insieme ottimo di variabili in grado di discriminare le classi.

Il problema di integrare il processo di selezione nel modello di classificazione è un problema noto in letteratura e sono state proposte molte soluzioni.

Considerato quindi un data set formato da N coppie $\{\mathbf{x}_i, y_i\} \quad i = 1, \ldots, N$ dove $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{Gi})$ è il vettore delle feature relativo al soggetto i e lo scalare y_i mi da la classe del soggetto i. É possibile calcolare la classe del soggetto i tramite un modello di regressione lineare, ossia possiamo cercare un vettore $\hat{\beta} = [\beta_1, \beta_2, \ldots, \beta_G]$ tale che $y = \hat{\beta}\mathbf{x}$. Una possibile soluzione a questo problema è il metodo basato sui minimi quadrati che cerca di trovare la soluzione del seguente problema di minimizzazione

$$\hat{\beta} = \arg\min_\beta \{E[y_i - \beta\mathbf{x}_i]\} = \arg\min_\beta \left\{ \frac{1}{N} \sum_{i=1}^{N} [y_i - \beta\mathbf{x}_i]^2 \right\}$$

Questo metodo però porta ad una soluzione in cui i pesi assegnati ai vari geni sono tutti non nulli, ossia si ha che $\hat{\beta} = (\beta_1, \ldots, \beta_i, \ldots, \beta_G) | \beta_i \neq 0 \ \ \forall i$, risulta allora impossibile selezionare.

Per ovviare a questo problema è possibile introdurre un parametro di regolarizzazione λ, ossia è possibile pensare di studiare un problema del tipo

$$\hat{\beta} = \arg\min_\beta [m(\beta, \mathbf{x}, y) + \lambda \times pen(\beta)]$$

In cui si definisce la funzione costo $m(\beta, \mathbf{x}, y) = ||y - \mathbf{x}\beta||_2^2$ e il termine $pen(\beta)$ è una funzione di penalizzazione che controlla la complessità del modello. Segue allora che

se $\lambda = 0$ ho un fit migliore dei dato ma un classificatore molto complesso, risulterebbe cioè una predizione non efficiente e soprattutto poco interpretabile

se $\lambda \to \infty$ risulteranno poche variabili usate per la classificazione (si avranno numerosi pesi $\beta_i = 0$, si parla quindi di soluzioni sparse).

Una possibile implementazione di questi metodi può essere il metodo *Lasso regression* [Tib-

shirani, 1996], in cui si definisce $pen(\beta) = ||\beta||_1$: la minimizzazione quindi è la seguente

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left[||y - \mathbf{x}\beta||_2^2 + \tau \, ||\beta||_1 \right]$$

Tale metodo promuove delle soluzioni sparse e quindi risulta particolarmente efficiente nella selezione di variabili, presenta tuttavia alcuni problemi tra cui il fatto che la selezione dei pesi β_i risente molto del rumore associato alle misure di espressione e il fatto che non si tiene conto dell'alta correlazione presente tra i geni che caratterizza i dati di espressione genica. Un'altra possibile implementazione di un metodo a penalizzazione è quella proposta da Hastie nel 2001 chiamata *ridge regression*, per questo tipo di implementazione viene posta $pen(\beta) = ||\beta||_2^2$ e si ha quindi un problema di minimizzazione del tipo

$$\hat{\beta}_{\text{RIDGE}} = \arg\min_{\beta} \left[||y - \mathbf{x}\beta||_2^2 + \epsilon \, ||\beta||_2^2 \right]$$

In questo caso il metodo garantisce soluzioni stabili e soprattutto tende a distribuire gli stessi pesi su variabili correlate, inoltre, specialmente in presenza di rumore, risulta che $\hat{\beta}$ contiene molti coefficienti bassi.

Un buon compromesso per integrare la selezione nel processo di classificazione dovrebbe tenere in considerazione sia le proprietà di selezione della *lasso regression* che le proprietà di gestione dei geni correlati della *ridge regression*. Si può definire allora una funzione costo che tenga conto di entrambe le proprietà definendo:

$$pen(\beta) \propto ||\beta||_1 , ||\beta||_2^2$$

si parla in questo caso di *elastic net* (Zou e Hastie, 2005) e la minimizzazione si formalizza nel seguente modo

$$\hat{\beta}_{\text{EN}} = \arg\min_{\beta} \left\{ ||y - \mathbf{x}\beta||_2^2 + \tau \left[||\beta||_1 + \epsilon \, ||\beta||_2^2 \right] \right\} \tag{3.3}$$

Si vede quindi che il parametro ϵ può essere usato per fissare il livello di selezione che si desidera, risulta infatti che

se ϵ diminuisce aumenta il peso di τ e di conseguenza diminuisce anche il numero delle variabili selezionate

se ϵ aumenta allora aumenta anche il numero di variabili selezionate altamente correlate

Sulla base dell'impianto teorico delle *elastic-net*, in [3] è stata proposta una variante al metodo che prevede una procedura su due stage, il primo atto a individuare un set minimo di geni in grado di eseguire una predizione precisa, il secondo per migliorare la stima di β.

Stage 1. Risolvo il problema di minimizzazione 3.3 fissando a priori $\epsilon = \epsilon_0 \sim 0$, calcolo quindi

$$(\tau_{opt}, \hat{\beta}) = \arg\min_{\beta,\tau} \left\{ ||y - \mathbf{x}\beta||_2^2 + \tau \left[||\beta||_1 + \epsilon \, ||\beta||_2^2 \right] \right\}$$

STAGE 2. Minimizzo la funzione costo

$$\left|\left|y - \tilde{\mathbf{x}}\tilde{\beta}\right|\right|_2^2 + \lambda \left|\left|\tilde{\beta}\right|\right|_2^2$$

dove $\tilde{\mathbf{x}}$ e $\tilde{\beta}$ sono rispettivamente la matrice di ingresso e il vettore dei pesi ristretti ai geni selezionati al passo 1.

Alla fine dei due step di analisi avrò quindi $(\tau_{opt}, \lambda_{opt})$ calcolati in maniera da minimizzare l'errore di validazione. Da tali valori sarà poi possibile ricavare il vettore dei pesi per un ϵ fissato, ossia fissando un livello di correlazioni sulle variabili.

3.4 Bootstrap

Se a livello teorico la valutazione delle performance di classificazione risulta semplice in pratica però tutta l'informazione che abbiamo riguardo ai soggetti è quella che siamo in grado di ricavare dal nostro data set, e vorremmo utilizzarla tutta per costruire un buon classificatore. Una volta però che un soggetto è stato utilizzato per istruire il classificatore non è lecito utilizzarlo per eseguire anche il test sulle performance del classificatore stesso.

Una prima soluzione al problema potrebbe essere quella di dividere il nostro data set iniziale in due insiemi, uno utilizzato per costruire il classificatore *training set* e un'altro, che chiameremo *test set*, usato solamente per testare il classificatore. In questo modo però l'informazione portata dai soggetti che finiscono nel test set va sprecata.

Per ovviare a questo problema si sono sviluppate delle strategie diverse per la valutazione delle performance di classificazione come ad esempio strategie di tipo *bootstrap*. lo schema di questo tipo di algoritmi è il seguente:

```
FOR (i in 1:B) :

        A. Campiona M soggetti (con ripetizione) dal DATA SET
             e crea il TRAINING SET i-esimo

        B. Definisci il TEST SET i-esimo come l'insieme degli
             elementi nel DATA SET originale non inclusi nel
             TRAINING SET i-esimo

        C. Applicando l'algoritmo di classificazione al
             TRAINING SET calcola le
             " regole di classificazione " --> RULES(i)

        D. Calcola l'errore medio su tutti i soggetti
             del TEST SET i-esimo --> ERR (i)

END
```

Infine l'errore di classificazione è dato dall'errore medio sui B errori medi calcolati all'interno del ciclo, cioè

$$errore = \frac{1}{B}\sum_{i=1}^{B} \mathrm{ERR}_i$$

3.5 Cross validazione

L'idea dei metodi di cross validazione è quella di dividere i dati in K sottoinsiemi e utilizzare un sottoinsieme alla volta come test set.

Considerato l'i-esimo set di dati come test set si usano i rimanenti $K-1$ come training set per il classificatore, si definiscono quindi le regole di classificazione trovate eliminando l'i-esimo set dai dati ($\textsc{rules}_{-i}$) ossia si calcola il modello di classificazione sull'insieme di dati ricavato escludendo l'i-esimo set. A partire da queste regole calcolo l'errore di classificazione sul soggetto j appartenente al set i ($\textsc{err}_j^{-i}$). L'errore varrà 0 se il soggetto j viene classificato correttamente e 1 se codificato in maniera errata.

Definisco quindi l'errore di cross validazione come

$$\textsc{cvError} = \frac{1}{K}\sum_{i=1}^{K} \frac{1}{\#\textsc{sets}_i} \sum_{j=1}^{\#\textsc{sets}_i} \textsc{err}_j^{-i}$$

dove $\#\textsc{sets}_i$ è il numero di soggetti che cadono nell'i-esimo set di dati. L'errore di cross validazione è quindi la media degli errori medi fatti per i K classificatori.

Capitolo 4

L'uso dell'INFORMAZIONE BIOLOGICA negli algoritmi di classificazione

I recenti sviluppi nell'ambito delle tecnologie di sequenziamento hanno permesso un notevole incremento delle conoscenze in ambito biologico. Tali conoscenze però risulterebbero inutilizzabili senza una organizzazione coerente.

Le nuove conoscenze che si hanno sui geni e sui loro prodotti infatti provengono da numerosi campi della biologia, la stessa informazione quindi potrebbe essere espressa in termini diversi da diversi studiosi a seconda del loro campo di appartenenza, risulta quindi evidente l'esigenza di un linguaggio comune.

Il *Gene Ontology Consortium* ha cercato di rispondere a questa esigenza proponendo appunto la *Gene Ontology* (GO), un'ontologia in continua espansione che raccoglie numerosissimi termini legati ai geni e ai loro prodotti.

Questo capitolo, dopo aver introdotto l'ontologia in senso generale, ha lo scopo di presentare nel dettaglio la GO descrivendo in particolare le regole che ogni termine deve rispettare e i legami tra i termini che formano la struttura dell'ontologia.

Il sequenziamento del genoma di molti organismi ha evidenziato il fatto che molte delle funzioni biologiche attivate dai geni sono condivise tra tutti gli organismi eucarioti. Questo comporta che conoscere il ruolo di un gene o di una proteina in un determinato organismo permette di trasferire questa conoscenza a tutti gli organismi che la condividono. L' obbiettivo del *Gene Ontology Consortium* è quello di costruire un vocabolario controllato, dinamico e che sia applicabile a tutti gli eucarioti. Il Gene Ontology Consortium è un progetto collaborativo nato con l'obbiettivo di unificare l'informazione su prodotti genici di vari database. Il progetto è iniziato nel 1998 dalla collaborazione di tre database contenenti informazioni su tre differenti organismi modello. FlyBase (*drosophila*, un moscerino della frutta), Saccharomyces Genome Database (lievito), e mouse genome database.

Da allora il consorzio è continuato a cresce includendo numerosi altri database di altri organismi.

Nella maggioranza dei casi infatti la nomenclatura con cui settori diversi della biologia si riferiscono agli stessi geni e ai loro prodotti rimane divergente. Il database go affronta questo problema permettendo di definire in modo corretto e non arbitrario i processi biologici cui un prodotto genico partecipa, le sue funzioni molecolari e la sua localizzazione cellulare.

L'idea di avere un linguaggio comune per descrivere i prodotti biologici permette di facilitare il mantenimento, l'aggiornamento e la gestione ottimale del database e inoltre fornisce una solida base su cui poggiare un archivio dinamico in grado di essere applicato a tutti gli eucarioti e capace di gestire in maniera ottimale il continuo aumento di informazioni su geni, proteine e loro ruolo nella cellula relativo ad una materia estremamente dinamica come è il mondo della biologia cellulare. Se inoltre il linguaggio è un linguaggio formale come è quello previsto da un ontologia è anche possibile pensare ad una gestione automatizzata della conoscenza.

4.1 Le tre categorie BP, MF, CC

Il database Gene Ontology è costituito da un ontologia di termini ben definiti che rappresentano le proprietà dei prodotti genici. L'ontologia è formata da tre categorie: *biological process* (BP), *molecular function* (MF) e *cellular component* (CC).

Biological process. si riferisce agli obbiettivi biologici a cui un gene, o un suo derivato, contribisce. Un processo biologico è portato a termine per mezzo di una o più funzioni molecolari ordinate, coinvolgendo trasformazioni di carattere chimico o fisico, Un esempio di termine specifico che si può trovare in BP è *cell growth and maintenance* o *trasduction*. Esempi di termini più più generali sono *pyramidine metabolism* o *cAMP biosynthesis*.

Molecular function. Si definisce molecular function ogni attività biochimica di un prodotto genico. Questa definizione si può anche applicare alle capacità potenziali che un prodotto genico può esprimere in un determinato contesto metabolico, senza tuttavia specificare dove viene svolta la funzione. Esempi di termini MF specifici sono *enzyme*,

transporter o *ligando*. Esempi di termini più generali possono essere invece *adenylate cyclase* o *Toll receptor ligand*.

CELLULAR COMPONENT. Si riferisce al luogo in cui un prodotto genico è attivo all'interno della cellula. Questi termini riflettono la nostra conoscenza sulla struttura della cellula eucariota. CC Esempi di termini GO di questa categoria sono *ribosome* o *proteosome*.

4.2 Il vocabolario controllato

Un termine GO si presenta nel modo seguente

```
id: GO:0016049
name: cell growth
namespace: biological_process
def: "The process in which a cell irreversibly increases
          in size over time by accretion and biosynthetic
          production of matter similar to that already present." [GOC:ai]
subset: goslim_generic
subset: goslim_plant
subset: gosubset_prok
synonym: "cell expansion" RELATED [ ]
synonym: "cellular growth" EXACT [ ]
synonym: "growth of cell" EXACT [ ]
is_a: GO:0009987 ! cellular process
is_a: GO:0040007 ! growth
relationship: part_of GO:0008361 ! regulation of cell size
```

possiede quindi un identificativo unico di 7 cifre con prefisso GO:. La parte numerica del GOID non ha nessuna relazione con la posizione del termine nell'ontologia. In *namespace* indica a quale delle tre categorie il termine appartiene.

Ad ogni termine è associato un campo *definition* che contiene una descrizione testuale di ciò che il termine rappresenta e un riferimento alla fonte dell'informazione su quel termine. I campi *subset* indicano l'appartenenza del termine ad alcune sotto-ontologie[1].

Il campo synonym invece permette di gestire eventuali sinonimi del termine, specificando in che modo ogni sinonimo si correla con il termine principale secondo vari livelli:

EXACT. indica l'esatta equivalenza tra il termine e il suo sinonimo, i due sono interscambiabili

BROAD. il sinonimo è più generale rispetto al termine

NARROW. il sinonimo è simile o più preciso rispetto al termine

RELATED. il sinonimo è in una certa relazione o copre in parte il significato del termine

[1]Dalla GO principale infatti sono state ricavate alcune sotto-ontologie relative a termini specifici, ad esempio PLANT SLIM o SCHIZOSACCHAROMYCES POMBE SLIM

Le associazioni tra il termine GO e i prodotti genici, dette annotazioni, sono disponibili in diversi database, a seconda del tipo di annotazione che si vuole utilizzare (e.g. geni, proteine). Per i dati ottenuti da microarray Affymetrix, è possibile ottenere le annotazioni GO relative ai trascritti interrogando il database NetAffx [4], che riporta le relazioni tra i termini GO e i trascritti interrogati dal microarray. Ogni termine GO può contenere uno o più trascritti annotati e i trascritti possono essere annotati su più termini GO.

4.3 L'ontologia

Una volta decritto il concetto di termine GO rimane da analizzare come sono legati i vari termini nell'ontologia. I termini GO rappresentano i nodi di una rete organizzata secondo un grafo diretto aciclico DAG. Tale struttura (DAG) differisce dalla classica struttura gerarchica in cui ogni nodo può avere un solo genitore, consente ad un nodo di avere più nodi padre. La struttura possiede due caratteristiche fondamentali per la gestione della conoscenza: è dinamica, perché l'ontologia cresce mano a mano che l'informazione si accumula, ed è flessibile dato che con una sola ontologia è possibile gestire le conoscenze che abbiamo su più organismi.

Come ogni termine deve sottostare a specifiche ben precise anche le relazioni tra i vari termini GO sono ben definite. Tali relazioni comprendono *is a* (*is a subtype of*), *part of*, *regulates*.

Vediamo brevemente le proprietà delle varie relazioni

is a. La relazione *is a* è la relazione più semplice: se si dice che, per due nodi A e B dell'ontologia, A is a B, si sta dicendo che il nodo A è un sottotipo del nodo B. Ad esempio mitotic cell cycle is a cell cycle oppure lyase activity is a catalytic activity.

A livello formale si ha che A is a B se e solo se data un'istanza a di A al tempo t allora a è un istanza anche di B al tempo t.

item*part of.* La relazione *part of* è usata per rappresentare le relazioni parti/tutto prensenti nella GO. Un arco di tipo part of può essere aggiunto tra due termini GO A e B se e solo se B esiste solo come parte di A e la presenza di B implica la presenza di A. Ad esempio replication fork è part of chromosome: ogni forca replicativa esiste solo come parte di qualche cromosoma, Mut solo alcuni cromosomi hanno una forca replicativa.

regulates. Un'altra relazione comune nel database è quella in cui un processo influisce direttamente sulla manifestazione di un'altro processo. L'obbiettivo della regolazione potrebbe essere un altro processo oppure una proprietà, come ad esempio le dimensioni della cellula o il pH. Analogamente alla relazione part of, la relazione *regulates* si usa solamente se, quando due processi A e B sono presenti contemporaneamente, uno regola sempre l'altro ma non vale il viceversa.

Ad esempio quando un cell cicle checkpoint è presente questo regola sempre il ciclo cellulare mentre il ciclo cellulare non è regolato solamente dal cell cycle checkpoints. Va detto poi che la regolazione di un processo non deve necessariamente far parte del processo stesso.

Consideriamo ora una relazione del tipo A is a B, in questo caso dirò che il termine A è più specifico rispetto al termine B. Generalizzando si ha che per qualsiasi relazione è possibile definire un grado di specificità tra i termini.

Si può allora pensare di ordinare i termini della GO in base a loro livello di significatività, per fare questo si deve definire un livello di significatività zero, è naturale considerare per questo la radice. Si tratta quindi di ordinare i nodi in base alla loro distanza dalla radice, emerge però un problema: in una struttura DAG ci sono più modi per collegare due nodi e quindi ci sono più modi per definire una distanza tra i due nodi. Per definire il livello di un nodo si dovrà decidere come scegliere tra percorsi alternativi.

In questo contesto definiremo il livello di un nodo come la sua distanza massima dal nodo radice, tale definizione infatti lega il livello più alto al grado di significatività maggiore.

4.4 Feng Tai e Wei Pan: LDA per gruppi di geni

Nel loro lavoro Feng Tai e Wei Pan hanno proposto la modifica del metodo PAM in modo di integrare le conoscenze sulle relazioni tra i geni nella classificazione e nella selezione dei geni.

L'idea è quella di considerare gruppi di geni suddivisi in base a conoscenze a priori secondo la loro appartenenza ad uno stesso pathway o più in generale rispetto ad una loro correlazione dal punto di vista biologico. L'ipotesi è quella che i geni appartenenti ad uno stesso gruppo siano correlati, ma geni appartenenti a gruppi diversi siano indipendenti tra loro.

Nel loro lavoro gli autori hanno quindi adattato la matrice di covarianza definita per il metodo PAM in modo da tener conto dell'indipendenza inter-gruppi, la matrice viene quindi definita in questo modo:

$$\tilde{\Sigma} = \alpha_1 \hat{\Sigma} + \alpha_2 \hat{\Sigma}^* + (1 - \alpha_1 - \alpha_2)\hat{\mathbf{D}}$$

con α_1 , $\alpha_2 \in [0, 1]$ parametri da determinare, in alternativa gli autori definiscono

$$\tilde{\Sigma} = \alpha \hat{\Sigma}^* + (1 - \alpha)\hat{\mathbf{D}} \quad , \ \alpha \in [0, 1]$$

Dove sono state utilizzate:

$\hat{\Sigma}$: matrice di covarianza dei campioni

$\hat{\mathbf{D}} = \text{diag}(\hat{\Sigma})$ matrice diagonale con le varianze dei campioni sulla diagonale

$\hat{\Sigma}^* = \text{diag}(\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_G)$ matrice diagonale a blocchi con $\hat{\Sigma}_i$ matrice di covarianza del gruppo i.

Chiameremo le analisi con i discriminanti definiti ora rispettivamente GRDA-1 e GRDA-2.

Dopo aver definito le matrici di covarianza vengono definiti anche due schemi di riduzione detti *shrinkage* basati sulle seguenti sogliature:

$$\hat{\mu}^*_{k(s)} = \text{sign}(\hat{\mu}^*_k)(|\hat{\mu}^*_k| - \lambda)_+ \quad , \quad \hat{\mu}^*_k = \tilde{\Sigma}^{-1}\hat{\mu}_k$$

tale schema però non è in grado di fare gene selection. La seconda soglia tende a mantenere o a rimuovere tutte le variabili o i geni in un gruppo. Invece di porre una soglia globale su tutte le $\hat{\mu}^*_{ik}$. In coppia con la seconda matrice definita prima, che assume l'indipendenza tra gruppi, il metodo è in grado di eseguire selezione di geni a livello di gruppo. In particolare si ha

$$\hat{\mu}^*_{kgs} = \hat{\mu}^*_{kg} \left(1 - \frac{\lambda \sqrt{p_g}}{\left|\left| \hat{\mu}^*_{kg} \right|\right|} \right)$$

dove

$\left|\left| \hat{\mu}^*_{kg} \right|\right|$ è la norma L_2 del vettore $\hat{\mu}^*_{ik}$

$p_g = |G_g|$ è la dimensione del gruppo g

Date due matrici di covarianza e due metodi per il calcolo delle soglie sui geni si hanno tre possibili metodi:

1. GRDA-1 con shrinkage method 1
2. GRDA-2 con shrinkage method 1
3. GRDA-2 con group shrinkage method

Dall'articolo emerge che il terzo metodo, che è stato confrontato con PAM, WPAM e SVM su dati di cancro, ha permesso di raggiungere performance di classificazione migliori. In più i geni selezionati hanno una interpretazione biologica in quanto vengono considerati gruppi di geni che appartengono allo stesso pathway biologico. Riportando i risultati sul breast cancer ad esempio si vede che vengono individuati con alta frequenza pathway significativi legati alla patologia (tabella 4.4).

PATHWAY ID	DESCRIPTION	FREQ
04010	MAPK signaling pathway	100
04360	Axon guidance	100
04060	Cytokine-cytokine receptor interaction	100
01430	Cell Communication	100
04080	Neuroactive ligand-receptor interaction	100
04730	Long-term depression	100
04020	Calcium signaling pathway	100
04510	Focal adhesion	99
04740	Olfactory transduction	99
02010	ABC transporters	97

Tabella 4.1: In tabella i 10 pathway (KEGG) selezionati con frequenza maggiore basati su 100 modelli calcolati con 10 ripetizioni di 10-fold cross validation.

4.5 Lottaz e Spang: classificare sulla GO

L'idea alla base del metodo proposto a da Lottaz e Spang [10] è quella di costruire un classificatore per ogni nodo della GO, ognuno di questi classificatori dipenderà dal livello di espressione genica dei geni relativi ai fenomeni biologici descritti dal nodo stesso.

Una volta costruito il grafo dei classificatori quindi viene applicata una soglia ai nodi GO eliminando i rami che sono meno probabilmente legati al fenotipo analizzato. I nodi rimanenti rappresentano i fenomeni biologici significativi per i campioni analizzati.

Il metodo, Structured Analysis of Microarrays (StAM) si compone essenzialmente di quattro passi:

1. creazione di un grafo con radice seguendo la struttura del grafo GO.

2. costruire i classificatori sui nodi foglia utilizzando approcci di classificazione classici

3. propagare i risultati ai nodi superiori fino ad arrivare alla radice

4. mettere una soglia al grafo dei classificatori in modo da determinare un piccolo insieme di funzioni significative

Nella creazione del grafo si tiene in considerazione il fatto che nella GO i geni possono essere annotati sia sui nodi foglia che sui nodi interni. Per questo, il grafo della GO viene arricchito con nuovi nodi foglia, in particolare per ogni nodo interno i con geni annotati in esso viene introdotto un nuovo nodo foglia i' al grafo con i come suo unico genitore. Quindi tutti i geni annotati in i andranno in i'.

Per rimuovere i nodi non significativi dai risultati vengono utilizzati due metodi: (1) Un nodo viene eliminato se ne lui ne i suoi successori hanno delle probe annotate. (2) Un nodo con un solo figlio è sostituito dal proprio figlio.

Ogni nodo foglia a conterrà quindi un set di nodi annotati, il classificatore corrispondente verrà costruito considerando solamente tali geni. Il classificatore utilizzato in questo lavoro è lo *shrunken centroid classificator* (Tibshirani et al., 2002).

Una volta che i classificatori sono stati costruiti sui nodi foglia, il metodo procede a trattare i nodi interni, in questa fase non verranno costruiti altri classificatori ma si procederà a combinare i risultati dei classificatori costruiti in precedenza. In particolare gli autori propongono una somma pesata dell'uscita della classificazione sui nodi figlio per propagare i risultati. I figli verranno pesati sulla base delle performance di classificazione.

Gli autori in particolare definiscono una misura di similarità d_i come

$$d_i = \frac{-2(1-\beta)}{|S_d|} \sum_{s \in S_d} \log(p_i^S) + \frac{-2\beta}{|S_c|} \sum_{s \in S_c} \log(1 - p_i^S)$$

dove S_c e S_d rappresentano rispettivamente i pazienti di controllo e i pazienti malati e p_i^s indica l'uscita del classificatore per il nodo i e il soggetto s. Si poi che d_i risulterà alto per i classificatori con basse performance mentre sarà basso per i classificatori buoni.

Successivamente, per eliminare i classificatori non informativi, si utilizza una soglia Δ e si eliminano i nodi con $\delta_i = 0$. Indicato con N_L l'insieme dei nodi foglia si calcolo

$$\delta_i = [\max_{j \in N_L}(d_j - d_i - \Delta)]^+$$

da tali pesi, indicato con $Ch(i)$ l'insieme dei figli del nodo i, si calcola per ogni nodo

$$w_{ij} = \frac{\delta_j}{\sum_{k \in Ch(i)} \delta_k}$$

$$p_i^s = \sum_{j \in Ch(i)} w_{ij} p_j^s$$

Una volta costruito il grafo dei classificatori i nodi meno significativi vengono eliminati basandosi su criteri di ridondanza della classificazione.

Dall'articolo emerge che il classificatore permette di raggiungere performance di classificazione paragonabili alle SVM ma è in grado di dare un'interpretazione dei risultati di tipo funzionale.

Capitolo 5

Il nostro metodo: Elimina e Classifica

Partendo da una panoramica sugli studi sul cancro il presente capitolo mira a evidenziare come la scienza sia sempre più consapevole della complessità dei meccanismi di regolazione genica.

L'espressione dei geni che vengono individuati dai metodi di analisi genomica classici è il risultato di una complessa rete di interazione che rappresenta la regolazione genica.

Tuttavia, tali geni non possono essere individuati se non pensando di individuare le interazioni che li legano, è chiara allora l'esigenza di sviluppare nuovi metodi di classificazione che permettano di tenere in considerazione le conoscenze sulla biologia che sono già a nostra disposizione. Tali metodi devono quindi essere in grado di integrare l'informazione biologica nei processi di classificazione.

Il nostro metodo si basa essenzialmente su due idee, la prima trae ispirazione dal metodo di Lottaz e Spang presentato nella sezione precedente e cioè prevede di definire dei classificatori associati ai nodi della GO, nel nostro caso però la costruzione dei classificatori si spinge anche nei nodi interni fino alla radice.

La seconda idea invece è quella di utilizzare il metodo *elim* introdotto da Adrian Alexa et al. in [11] opportunamente modificato. Tale metodo visita i nodi della GO con una strategia bottom-up. Inizia quindi a processare il grafo della GO dai nodi foglia spostandosi via via verso i nodi genitore, muovendosi dai nodi più specifici a quelli meno specifici.

Dato che i nodi dello stesso livello non condividono relazioni tra loro, questi possono essere investigati in maniera indipendente. La strategia bottom-up assicura poi che nel momento in cui ci si trova ad analizzare un certo nodo, tutti i suoi nodi figli sono già stati analizzati.

Nel momento in cui viene processato un nodo u, allora decido se il modello di classificazione costruito ha buone performance di classificazione, in questo caso allora andrò a selezionare il nodo (e di conseguenza i suoi geni) come significativo e provvederò ad eliminare il set di geni da tutti gli antenati di u.

Con questo accorgimento è chiaro che riuscirò a selezionare i nodi significativi più specifici che sono presenti nel grafo. Si supponga infatti che un nodo contenga lo stesso set di geni di uno dei suoi figli, in questo caso i metodi classici pesano allo stesso modo i due termini. Andando ad eliminare i geni dal nodo a livello più alto invece avrò che il nodo più significativo sarà il nodo figlio che è quello più specifico e quindi più utile.

Nel nostro caso la decisione di considerare il modello di di classificazione di un determinato nodo significativo o meno viene fatta considerando una soglia sull'MCC del classificatore calcolato sui geni appartenenti ad un nodo. Dato un classificatore si definisce MCC (*Matthews correlation coefficient*) l'indice calcolata a partire dalla matrice di confusione, come rapporto

$$\text{MCC}(\hat{y}|y) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

dove con $\hat{y}$ si indicano le predizioni stimate e con y le predizioni vere. L'MCC può assumere valori da -1 a 1 dove un MCC $= 1$ indica predizione perfetta, 0 indica predizione casuale e -1 indica predizione inversa.

Definiamo quindi un nodo i, e l'insieme dei geni in esso annotati $G(i)$, a questo viene associato un classificatore $l_1 l_2$, che indichiamo con C_i, formato considerando solo i geni appartenenti a i. Su tale classificatore viene calcolato l'MCC_i: se MCC_i è maggiore di una certa soglia si considera il nodo selezionato e si procede ad eliminare il set di geni $G(i)$ da tutti i nodi antenati di i.

Vediamo ora nel dettaglio come lavora il metodo. I dati in ingresso sono la matrice dei

valori di espressione genica $\mathbf{D}$, e il vettore delle label $\mathbf{Y}$

$$\mathbf{D} = \begin{bmatrix} g_{11} & g_{21} & \cdots & g_{i1} & \cdots & g_{G1} \\ g_{12} & g_{22} & \cdots & g_{i2} & \cdots & g_{G2} \\ g_{13} & g_{23} & \cdots & g_{i3} & \cdots & g_{G3} \\ \vdots & \vdots & \ldots & g_{ij} & \vdots & \vdots \\ g_{1M} & g_{2M} & \cdots & g_{iM} & \cdots & g_{GM} \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_M \end{bmatrix}$$

dove con g_{ij} è indicato il livello di espressione del gene $i \in \{1, G\}$ nel soggetto $j \in \{1, \ldots, M\}$ e con $y_i \in \{0, 1\}$ si indica la classe di appartenenza del soggetto i.

Come prima fase il metodo crea B coppie di insiemi *test set*, *training set*, si hanno allora

$$\left(\text{TR}^b, \text{TS}^b\right) \qquad b = 1, \ldots, B$$

dove

$$\begin{aligned} &\text{TR}^b = [tr_1, tr_2, \ldots, tr_M] \quad \text{con} \quad tr_i \in \mathcal{U}(1, M) \\ &\text{TS}^b = [1, 2, 3, \ldots, M] - \text{TR} \end{aligned}$$

L'algoritmo quindi per ogni coppia $\left(\text{TR}^b, \text{TS}^b\right)$ definisce

$$\begin{aligned} \mathbf{D}^{TR^b} \in \mathcal{M}(M, G) \quad &: \quad \mathbf{D}^{TR^b}_{ij} = \mathbf{D}_{\text{TR}^b_i,\ j} \\ \mathbf{D}^{TS^b} \in \mathcal{M}(M, G) \quad &: \quad \mathbf{D}^{TS^b}_{ij} = \mathbf{D}_{\text{TR}^b_i\ ,\ j} \end{aligned}$$

e

$$\begin{aligned} \mathbf{Y}^{TR^b} \in \mathcal{M}(M, 1) \quad &: \quad \mathbf{Y}^{TR^b}_i = \mathbf{Y}_{\text{TR}^b_i} \\ \mathbf{Y}^{TS^b} \in \mathcal{M}(M, 1) \quad &: \quad \mathbf{Y}^{TS^b}_i = \mathbf{Y}_{\text{TS}^b_i} \end{aligned}$$

Posto allora $i =$ MASSIMO LIVELLO IN DAG, l'algoritmo , per ogni coppia $\left(\text{TR}^b, \text{TS}^b\right)$ definita, procede in questo modo

1. SELEZIONE DEI NODI AL LIVELLO i. Si definisce l'insieme $\mathbf{N}_i = \{n_1, n_2, \ldots, n_l\}$ contenente tutti i nodi del grafo GO con livello uguale ad i e con un numero di annotazioni maggiore di 5.

2. COSTRUZIONE DEL CLASSIFICATORE E METODO *elim*. Per ogni nodo $n \in \mathbf{N}_i$ si costruisce un classificatore $l_1 l_2$ considerando solamente i valori di espressione genica dei geni annotati in n, si risolve quindi il problema

$$\hat{\beta}_n = \arg\min_{\beta} \left\{ \left|\left| \mathbf{Y}^{TR^b}|_{G(n)} - \mathbf{D}^{TR^b}|_{G(n)} \times \beta \right|\right|_2^2 + \tau \left[||\beta||_1 + \epsilon \, ||\beta||_2^2 \right] \right\}$$

dove con $G(n)$ si indicano i geni annotati nel nodo n e con $\mathbf{D}^{TR^b}|_{G(n)}$ si indica l'operazione di restringere i geni della matrice $\mathbf{D}^{TR^b}$ ai geni appartenenti ad n. I parametri ottimi per il classificatore l_1l_2 vengono calcolati mediante una *k-cross-validation* con $k = 5$ (vedi 3.5).

Una volta calcolato il $\hat{\beta}_n$ si calcola il valore di MCC del classificatore sui dati del test set, si calcola quindi

$$\text{MCC}\left(\hat{\mathbf{Y}} = \mathbf{D}^{TS^b} \times \hat{\beta}_n \mid \mathbf{Y}^{TS^b}\right)$$

se tale valore supera una certa soglia prefissata th_{mcc} allora il nodo si considera selezionato e si procede ad eliminare i nodi selezionati in n dai suoi antenati (procedura *elim*).

Alla fine di questo secondo passaggio si hanno un insieme di vettori dei pesi β_i^* formato dai vettori dei pesi dei nodi selezionati al livello i.

3. AGGIORNAMENTO DI i. Una volta completati i due step precedenti si passa ad un livello inferiore $i = i - 1$ e si ricomincia dal punto 1.

Il valore di th_{mcc} viene calcolato in una fase inizale del metodo come 95° percentile della distribuzione degli MCC calcolata sui classificatori definiti applicando il metodo stesso senza però eseguire la procedura *elim*.

Quando si arriva al livello 0 l'algoritmo conclude.

Alla fine si avranno, come uscita per ogni coppia $\left(\text{TR}^b, \text{TS}^b\right)$, un insieme di vettori β_i^* per ogni livello della GO.

Dato quindi un nuovo soggetto $\mathbf{s} = (g_{1s}, g_{2s}, \ldots, g_{Gs})$ da classificare la predizione verrà calcolata come

$$\hat{y_s} = \frac{1}{\#\text{NS}} \sum_{i=1}^{\text{ML}} \left(\sum_{\hat{\beta}_j \in \beta_i^*} \mathbf{s} \times \hat{\beta}_j \right) \tag{5.1}$$

Dove con #NS si è indicato il numero di nodi selezionati e con ML si indica il massimo livello della GO. La predizione quindi sarà la media pesata delle predizioni fatte sui nodi selezionati, si parla in questo caso di *classificatore aggregato*. La predizione di nuovi soggetti, diversamente da quanto avviene per i metodi classici, viene fatta partizionando il problema originale in sottoproblemi di dimensione più piccola e raggruppando i risultati.

Nel dettaglio ad ogni nodo selezionato viene associato un modello di classificazione definito da $\hat{\beta}_j$, tale modello permette di calcolare la predizione $\hat{Y}_j$ per ogni nodo selezionato come $\hat{Y}_j = \hat{\beta}_j \times \mathbf{s}$. Si noti che $-1 \leq \hat{Y}_j \leq 1$, il valore assoluto di $\hat{Y}_j$ quindi può essere interpretato come il grado di certezza del classificatore j. Un classificatore aggregato pesato, come indicato in 5.1, fa la media delle predizioni dei vari classificatori e classifica nel seguente modo: se $\hat{y_s} < 0 \to -1$, mentre se $\hat{y_s} > 0 \to 1$. In questo modo i classificatori che classificano "meglio" pesano di più nella decisione finale. Un'altra soluzione è quella di costruire un classificatore aggregato non pesato dove, anziché considerare la media delle variabili $\hat{Y}_j$, si considera la

media delle variabili $\hat{Y}_j^*$ definite nel modo seguente

$$\hat{Y}_j^* = \begin{cases} 1 \quad \text{se} \quad Y_j > 0 \\ -1 \quad \text{se} \quad Y_j < 0 \end{cases}$$

in questo modo ogni classificatore ha lo stesso peso nella decisione finale.

Il metodo qui descritto è stato sviluppato in python sfruttando il pacchetto `l1l2py` per il classificatore $l_1 l_2$.

Capitolo 6

Risultati

Il metodo è stato applicato su dati di espressione genica e i risultati ottenuti sono stati valutati in termini di performance di classificazione e riproducibilità delle informazioni biologiche selezionate. In particolare, è stata considerata sia l'accuratezza nella predizione delle classi, calcolata in termini di MCC, sia la frequenza dei termini GO selezionati confrontando i risultati ottenuti su diversi dataset di soggetti affetti dalla stessa patologia. Infine, l'interpretabilità dei risultati è stata esplorata sui termini GO selezionati e valutando se questi possono essere comparabili tra i dataset e coerenti con il tipo di patologia considerata. Per l'applicazione ai dati, il metodo è stato implementato sviluppando un algoritmo in linguaggio Python facilmente parallelizzabile per poter eseguire il metodo contemporaneamente su più dataset diversi. I risultati forniti in output dall'algoritmo possono essere riassunti nei seguenti punti:

a. Lista dei nodi selezionati dal classificatore

b. I geni selezionati dal classificatore l1l2 per ogni nodo selezionato

c. I parametri stimati per ogni modello di classificazione stimato su ogni nodo

d. Le predizioni ottenute dai classificatori associati ai nodi con il risultato finale del classificatore aggregato

6.1 I Dati

Il metodo descritto in 5 è stato applicato su tre dataset pubblici di breast cancer disponibili nel database Gene Expression Omnibus [12] con i seguenti codici identificativi: GSE2990, GSE3494 and GSE7390. I dataset contengono dati di espressione estratti da diversi soggetti utilizzando microarray Affymetrix U133 GenechipsTM (HG-U133A) e le classi sono state definite in base al fattore clinico che indica lo stato del recettore dell'estrogeno (ER^+ e ER^-). Le popolazioni di soggetti monitorati sui tre dataset sono quindi così distribuite: GSE2990 con 179 soggetti (145 ER^+, 34 ER^-), GSE3494 con 155 soggetti (131 ER^+ e 24 ER^-) e GSE7390 con 152 soggetti (103 ER^+, 49 ER^-). Per ogni soggetto si conoscono i livelli di espressione di 22207 feature (probe sets). Tali dati sono stati classificati con il metodo proposto considerando due

delle tre categorie GO: BP e MF. Si avranno quindi sei casi: 1) GSE 2990-BP, 2) GSE2990-MF, 3) GSE3494-BP 4) GSE3494-MF, 5)GSE7390-BP e 6) GSE7390-MF.

La struttura del grafo GO è stata estratta dalle informazioni riportate in `gene_ontology.obo` versione 1.2 (09/16/2011), disponibile sul sito `http://www.geneontology.org`. Il file con le informazioni sulle annotazioni invece è stato scaricato all'indirizzo `http://www.ebi.ac.uk/GOA/` il 11/06/2011.

6.2 Valutazione delle performance di classificazione

La valutazione delle performance in termini di classificazione è stata eseguita applicando il metodo su 10 coppie training set - test set (B = 10) utilizzando l'approccio bootstrap 3.4. La soglia th_{mcc} viene calcolata come 95° percentile della distribuzione dei valori di MCC sui nodi GO calcolati secondo il metodo descritto nel capitolo 5 senza l'applicazione del metodo elim.

In particolare per ogni coppia (TR, TS) si è calcolato l'MCC del classificatore aggregato, costruito come descritto in 5.

Le coppie create inoltre sono state utilizzate per eseguire una classificazione utilizzando un approccio standard[1], i cui risultati verranno utilizzati come riferimento. In tabella 6.1 e figura 6.1 è possibile osservare i risultati sull'accuratezza di predizione in termini di MCC nelle 10 coppie. Si può notare che, in media, le performance di classificazione aumentano sistematicamente per il classificatore aggregato basato sulla GO rispetto alla classificazione standard. Tali differenze risultano statisticamente significative confrontando i valori di MCC con un test di Wilcoxon, ottenendo p-value minori di 0.021.

6.3 Valutazione della stabilità delle liste dei nodi GO

Oltre al classificatore aggregato, l'algoritmo che implementa il metodo proposto fornisce, per ogni coppia (TR, TS), anche la lista di termini GO che sono stati selezionati applicando la soglia th_{mcc} sull'MCC ottenuto dal classificatore stimato in ogni nodo. In particolare, considerando le analisi effettuate per le 10 coppie (TR, TS), per valutare la riproducibilità dei risultati è stata considerata, per ogni dataset, la lista di termini GO che venivano selezionati con una percentuale sulla frequenza maggiore o uguale al 70% nelle 10 liste di termini generate. Nella tabella 6.2 è riportata, per ognuno dei tre DATASET, la percentuale dei termini con frequenza maggiore a 70% ottenuti dal metodo sulle due categorie GO considerate. Tali risultati sono stati confrontati con l'applicazione standard del metodo di classificazione. In quest'ultimo caso, tuttavia, si hanno a disposizione liste di geni anziché di termini GO. Per interpretare tali i risultati in termini di processi biologici e funzioni molecolari, un approccio molto comunemente usato in letteratura prevede di applicare un'analisi a posteriori alle liste di geni utilizzando le annotazioni GO e applicando un test statistico (Fisher's Exact Test) che permette di stabilire quanto le annotazioni appartenenti ad un termine GO caratterizzano la lista dei geni selezionati, assegnando un p-value ad ogni termine. Dalle 10 liste di geni

[1] senza integrazione dell'informazione biologica

ottenute dal metodo standard sono state quindi ottenute le corrispondenti liste di termini GO selezionati ponendo una soglia sul p-value pari a 0.01. Analizzando la tabella 6.2 (riquadro superiore sinistro) e confrontando con l'analisi standard è possibile notare che, in proporzione, il metodo di integrazione proposto è in grado di selezionare, nella maggior parte dei casi, un numero superiore di termini GO che risultano essere stabili nelle 10 coppie osservate.

Poiché i tre dataset considerati riguardano lo stesso caso clinico, è stata calcolata la proporzione di termini GO in comune in due o tutti e tre i dataset (riquadro inferiore tabella 6.2 e diagrammi di Venn riportati nelle figure 6.3). Come si può notare, i termini selezionati dal metodo di integrazione usato permettono di conservare in modo efficiente l'informazione biologica nei tre dataset.

Inoltre, osservando quali sono i termini che si ritrovano nei tre dataset (tabella 6.3) si vede che questi hanno particolare rilevanza per il caso patologico analizzato: per la categoria BP infatti si ritrovano i termini *regulation of trascription*, *cell cycle process* oppure *cell differentiation*, mentre per MF ci sono termini in comune legati allo stress cellulare come *oxidoreductase activity*. Cercando di approfondire in maggior dettaglio la consistenza dei risultati dal punto di vista biologico, i termini selezionati nei tre dataset sono stati raggruppati in macro-gruppi di processi biologici e funzioni molecolari in modo da valutare i principali meccanismi biologici che caratterizzano i risultati ottenuti. A tal scopo, è stato effettuato un clustering gerarchico agglomerativo (average linkage) sui termini GO selezionati utilizzando una distanza basata sulla matrice di similarità di Resnik, definita come

$$\text{SIM}_{Res}(c_1, c_2) = \max_{c \in \text{LCA}(c_1, 2_2)} \text{IC}(c)$$

Dove con $\text{LCA}(c_1, c_2)$ si definisce il primo antenato comune condiviso dai nodi c_1 e c_2, mentre IC indica l'*information content* della GO definito come

$$\text{IC}(c) = -\log\left[\frac{freq(c)}{freq(\text{ROOT})}\right]$$

dove $freq(c)$ è il numero di occorrenze delle annotazioni nel nodo c e in tutti i suoi nodi discendenti

$$\text{freq}(c) = |\text{annot}(c)| + \sum \text{freq}_{t \in \text{nodi-figlio}(c)}(t)$$

I risultati del clustering sono riportati, per quanto riguarda BP, nelle tabelle 6.4, 6.5 e 6.6 e per MF in 6.7, 6.8 e 6.9. Si può vedere come i termini individuati dall'intersezione dei tre dataset definiscano gruppi ben distinti che si mantengano nei tre dataset, in generale si possono definire i seguenti macro-gruppi:

REGOLAZIONE. In questo gruppo cadono i processi *regulation of transcription,* DNA-*dependent*, e *positive regulation of cellular process*, inoltre per il GSE 3494 e per il GSE 7390 tali termini sono legati ad esempio alla produzione di citochine o di interluchine che sono in grado di indurre ad esempio crescita e differenziazione cellulare. Si noti in oltre come nel dataset GSE3494 sia legato ai termini sopra indicati anche il termine *negative regulation of apoptosi*, che è strettamente legato al cancro.

Crescita/Sviluppo cellulare. In questo gruppo si possono trovare *developmental process involved in reproduction*, *anatomical structure morphogenesis*, *organ morphogenesis*, *cell differentiation*, *developmental process*, *organ development*, *system development*, *anatomical structure development* e *cellular developmental process*.

Risposta a stimoli. In questo gruppo si possono far ricadere i termini *response to stress*, *response to organic substance* e *response to stimulus*, nei vari dataset inoltre tali termini vengono legati anche a termini come *response to metal ion* e *response to oxygen levels* che riguardano la risposta della cellula a stress cellulari caratteristici per il cancro.

Metabolismo cellulare. I rimanenti processi tra quelli elencati in tabella 6.3 possono essere raccolti in un gruppo i cui termini sono inerenti alla produzione di sostanze.

I risultati del processo di clustering per quanto riguarda MF è riportato nelle tabelle 6.7, 6.8 e 6.9 e evidenzia l'esistenza di due macro-gruppi stabili legato al concetto

binding. legati a questo concetto possiamo ad esempio trovare termini legati al *metal ion binding* come può essere *transition metal ion binding*, in accordo anche con i risultati per BP.

Ossidoriduzione. Dove possiamo trovare termini come *oxidoreductase activity* che conferma la validità dei processi biologici selezionati, con particolare riferimento al macro-gruppo *risposta a stimoli*.

Si vede allora che integrare l'informazione contenuta nella GO ai metodi di classificazione mi permette di descrivere una patologia in base ai termini generali della GO e quindi in un ottica di tipo funzionale, questo è in netto in contrasto con i metodi di classificazione standard che danno in uscita solamente liste di geni difficilmente interpretabili, ad esempio, in ambito clinico.

Il confronto dei risultati ottenuti nei tre dataset mostrano poi che il metodo risulta promettente per la ricerca di biomarcatori su set di geni caratterizzati da funzionalità simili e coinvolti in processi biologici simili.

	GSE2990			GSE3494			GSE7390		
REP.	ST.	BP	MF	ST.	BP	MF	ST.	BP	MF
r_1	0,386	0,606	0,542	0,462	0,580	0,609	0,623	0,847	0,802
r_2	0,669	0,495	0,576	0,424	0,528	0,528	0,599	0,790	0,790
r_3	0,534	0,620	0,620	0,453	0,480	0,480	0,744	0,841	0,841
r_4	0,359	0,929	0,866	0,462	0,675	0,609	0,552	0,650	0,685
r_5	0,244	0,722	0,671	0,303	0,437	0,406	0,460	0,904	0,904
r_6	0,422	0,760	0,760	0,280	0,541	0,657	0,660	0,623	0,660
r_7	0,576	0,847	0,781	0,388	0,553	0,580	0,847	0,847	0,847
r_8	0,300	0,433	0,485	0,341	0,428	0,428	0,617	0,784	0,784
r_9	0,386	0,669	0,722	0,462	0,553	0,528	0,701	0,744	0,701
r_{10}	0,577	0,765	0,765	0,382	0,453	0,428	0,802	0,950	0,950
MEDIA	0,445	0,685	0,679	0.396	0.523	0.525	0.661	0.798	0.796
SD	±0.137	±0.152	±0.121	±0.069	±0.076	±0.088	±0.117	±0.104	±0.094

Tabella 6.1: Confronto tra media e deviazione standard dell'MCC relativa all'applicazione del metodo di classificazione standard e del metodo basato sulla GO.

	ANALISI ST.		METODO TESI	
	BP	MF	BP	MF
GSE2990	1,6 %	4,0%	1,9%	1,6%
GSE3494	0,3%	0,4%	3,1%	2,4%
GSE7390	5,1%	6,0%	6,3%	6,9%
CONDIVISI 2^+	4,5%	12,5%	34,0%	26,0%

Tabella 6.2: Tabella che confronta la percentuale di nodi selezionati per l'analisi standard e per il nostro metodo, l'ultima riga inoltre evidenzia la percentuale di nodi condivisa nei tre dataset. Si può vedere come, con il metodo proposto, si riescano a stabilizzare le liste di nodi selezionati.

Tabella 6.4: Tabella relativa ai cluster di geni significativi per il dataset GSE2990 analizzato per BP

GOID	GOTERM	[GSE2990] CLUSTER
GO:0001932	regulation of protein phosphorylation (8)	1
GO:0006355	regulation of transcription, DNA-dependent (7)	1
GO:0009966	regulation of signal transduction (4)	1
GO:0031328	positive regulation of cellular biosynthetic process (6)	1
GO:0048522	positive regulation of cellular process (4)	1
GO:0048523	negative regulation of cellular process(4)	1
GO:0050789	regulation of biological process (2)	1
GO:0051726	regulation of cell cycle (4)	1
GO:0003006	developmental process involved in reproduction (2)	2
GO:0009653	anatomical structure morphogenesis (2)	2
GO:0009887	organ morphogenesis (3)	2
GO:0030154	cell differentiation (3)	2
GO:0032502	developmental process (1)	2
GO:0046530	photoreceptor cell differentiation (5)	2
GO:0048513	organ development (3)	2
GO:0048610	cellular process involved in reproduction (2)	2
GO:0048731	system development (3)	2
GO:0048856	anatomical structure development (2)	2
GO:0048869	cellular developmental process (2)	2
GO:0050877	neurological system process (3)	2
GO:0006575	cellular modified amino acid metabolic process (7)	3
GO:0006732	coenzyme metabolic process (4)	3
GO:0008152	metabolic process (1)	3
GO:0009987	cellular process (1)	3
GO:0016568	chromatin modification (7)	3
GO:0022402	cell cycle process (2)	3
GO:0042423	catecholamine biosynthetic process (8)	3
GO:0044237	cellular metabolic process (2)	3
GO:0044242	cellular lipid catabolic process (5)	3
GO:0044248	cellular catabolic process (3)	3
GO:0044249	cellular biosynthetic process (3)	3
GO:0046520	sphingoid biosynthetic process (8)	3

GO:0006950	response to stress (2)	4
GO:0006979	response to oxidative stress (3)	4
GO:0007169	transmemb. receptor protein tyrosine kinase sig. path. (7)	4
GO:0009605	response to external stimulus (2)	4
GO:0010033	response to organic substance (3)	4
GO:0010038	response to metal ion (4)	4
GO:0033554	cellular response to stress (3)	4
GO:0042221	response to chemical stimulus (2)	4
GO:0050896	response to stimulus (1)	4
GO:0070482	response to oxygen levels (3)	4

Tabella 6.5: Tabella relativa ai cluster di geni significativi per il dataset GSE3494 analizzato per BP

GOID	GOTERM	[GSE3494] CLUSTER
GO:0001816	cytokine production (2)	1
GO:0032501	multicellular organismal process (1)	1
GO:0050877	neurological system process (3)	1
GO:0003008	system process (2)	1
GO:0042089	cytokine biosynthetic process (5)	1
GO:0003006	developmental process involved in reproduction (2)	2
GO:0009653	anatomical structure morphogenesis (2)	2
GO:0009887	organ morphogenesis (3)	2
GO:0030154	cell differentiation (3)	2
GO:0032502	developmental process (1)	2
GO:0048513	organ development (3)	2
GO:0048731	system development (3)	2
GO:0048856	anatomical structure development (2)	2
GO:0048869	cellular developmental process (2)	2
GO:0007399	nervous system development (4)	2
GO:0007409	axonogenesis (7)	2
GO:0022008	neurogenesis (4)	2
GO:0048485	sympathetic nervous system development (4)	2
GO:0048812	neuron projection morphogenesis (6)	2
GO:0006351	transcription, DNA-dependent (8)	3
GO:0006575	cellular modified amino acid metabolic process (7)	3
GO:0008152	metabolic process (1)	3

GO:0043170	macromolecule metabolic process (2)	3
GO:0044237	cellular metabolic process (2)	3
GO:0044248	cellular catabolic process (3)	3
GO:0044249	cellular biosynthetic process (3)	3
GO:0044260	cellular macromolecule metabolic process (3)	3
GO:0006520	cellular amino acid metabolic process (6)	3
GO:0008652	cellular amino acid biosynthetic process (7)	3
GO:0009058	biosynthetic process (2)	3
GO:0009059	macromolecule biosynthetic process (3)	3
GO:0034645	cellular macromolecule biosynthetic process (4)	3
GO:0044106	cellular amine metabolic process (4)	3
GO:0044238	primary metabolic process (2)	3
GO:0006355	regulation of transcription, DNA-dependent (7)	4
GO:0010557	positive regulation of macromolecule biosynthetic proc. (6)	4
GO:0031328	positive regulation of cellular biosynthetic process (6)	4
GO:0042127	regulation of cell proliferation (4)	4
GO:0048522	positive regulation of cellular process (4)	4
GO:0048523	negative regulation of cellular process (4)	4
GO:0050789	regulation of biological process (2)	4
GO:0006357	regulation of transcription from RNA polym. II prom. (8)	4
GO:0031326	regulation of cellular biosynthetic process (5)	4
GO:0032673	regulation of interleukin-4 production (5)	4
GO:0042035	regulation of cytokine biosynthetic process (6)	4
GO:0043066	negative regulation of apoptosis (7)	4
GO:0045893	positive regulation of transcription, DNA-dependent (8)	4
GO:0045944	positive regulat. of transcr. from RNA polym. II prom. (9)	4
GO:0051252	regulation of RNA metabolic process (6)	4
GO:0006950	response to stress (2)	5
GO:0010033	response to organic substance (3)	5
GO:0050896	response to stimulus (1)	5
GO:0006952	defense response (3)	5
GO:0035556	intracellular signal transduction (5)	5
GO:0009987	cellular process (1)	6
GO:0022402	cell cycle process (2)	7
GO:0051704	multi-organism process (1)	8

Tabella 6.6: Tabella relativa ai cluster di geni significativi per il dataset GSE7390 analizzato per BP

GOID	GOTERM	[GSE7390] CLUSTER
GO:0003006	developmental process involved in reproduction (2)	1
GO:0000003	reproduction (1)	1
GO:0007548	sex differentiation (3)	1
GO:0019953	sexual reproduction (2)	1
GO:0009653	anatomical structure morphogenesis (2)	2
GO:0009887	organ morphogenesis (3)	2
GO:0030154	cell differentiation (3)	2
GO:0032502	developmental process (1)	2
GO:0048513	organ development (3)	2
GO:0048731	system development (3)	2
GO:0048856	anatomical structure development (2)	2
GO:0048869	cellular developmental process (2)	2
GO:0001501	skeletal system development (4)	2
GO:0009952	anterior/posterior pattern specification (4)	2
GO:0031099	regeneration (3)	2
GO:0035239	tube morphogenesis (3)	2
GO:0035282	segmentation (4)	2
GO:0048568	embryonic organ development (4)	2
GO:0042127	regulation of cell proliferation (4)	3
GO:0048523	negative regulation of cellular process (4)	3
GO:0050794	regulation of cellular process (3)	3
GO:0048519	negative regulation of biological process (3)	3
GO:0001558	regulation of cell growth (4)	3
GO:0001816	cytokine production (2)	4
GO:0032501	multicellular organismal process (1)	4
GO:0001819	positive regulation of cytokine production (5)	5
GO:0048522	positive regulation of cellular process (4)	5
GO:0001817	regulation of cytokine production (4)	5
GO:0051239	regulation of multicellular organismal process (3)	5
GO:0002376	immune system process (1)	6

GO:0006351	transcription, DNA-dependent (8)	7
GO:0006464	protein modification process (5)	7
GO:0006575	cellular modified amino acid metabolic process (7)	7
GO:0008152	metabolic process (1)	7
GO:0042423	catecholamine biosynthetic process (8)	7
GO:0043170	macromolecule metabolic process (2)	7
GO:0044237	cellular metabolic process (2)	7
GO:0044249	cellular biosynthetic process (3)	7
GO:0044260	cellular macromolecule metabolic process (3)	7
GO:0009058	biosynthetic process (2)	7
GO:0009059	macromolecule biosynthetic process (3)	7
GO:0032787	monocarboxylic acid metabolic process (6)	7
GO:0034645	cellular macromolecule biosynthetic process (4)	7
GO:0010467	gene expression (3)	7
GO:0016310	phosphorylation (5)	7
GO:0032774	RNA biosynthetic process (7)	7
GO:0034961	cellular macromolecule biosynthetic process (4)	7
GO:0043284	macromolecule biosynthetic process (3)	7
GO:0006355	regulation of transcription, DNA-dependent (7)	8
GO:0006357	regulation of transcript. from RNA polymerase II prom. (8)	8
GO:0010468	regulation of gene expression (5)	8
GO:0010556	regulation of macromolecule biosynthetic process (5)	8
GO:0031323	regulation of cellular metabolic process (4)	8
GO:0031326	regulation of cellular biosynthetic process (5)	8
GO:0009889	regulation of biosynthetic process (4)	8
GO:0019222	regulation of metabolic process (3)	8
GO:0032318	regulation of Ras GTPase activity (11)	8
GO:0043393	regulation of protein binding (4)	8
GO:0046578	regulation of Ras protein signal transduction (6)	8
GO:0050790	regulation of catalytic activity (4)	8
GO:0060255	regulation of macromolecule metabolic process (4)	8
GO:0006810	transport (2)	9
GO:0008104	protein localization (3)	9
GO:0033036	macromolecule localization (2)	9
GO:0051179	localization (1)	9
GO:0051641	cellular localization (2)	9
GO:0006950	response to stress (2)	10
GO:0007165	signal transduction (4)	10
GO:0007169	transm. receptor protein tyrosine kinase sign. path. (7)	10

GO:0010033	response to organic substance (3)	10
GO:0010038	response to metal ion (4)	10
GO:0042221	response to chemical stimulus (2)	10
GO:0050896	response to stimulus (1)	10
GO:0006952	defense response (3)	10
GO:0009725	response to hormone stimulus (4)	10
GO:0007265	Ras protein signal transduction (7)	10
GO:0006996	organelle organization (4)	11
GO:0022402	cell cycle process (2)	12
GO:0007049	cell cycle (2)	12
GO:0007155	cell adhesion (2)	13
GO:0031589	cell-substrate adhesion (3)	13
GO:0009987	cellular process (1)	14
GO:0050793	regulation of developmental process (3)	15

Tabella 6.7: Tabella relativa ai cluster di geni significativi per il dataset GSE2990 analizzato per MF

GOID	GOTERM	[GSE2990] CLUSTER
GO:0004888	transmembrane signaling receptor activity (4)	1
GO:0005102	receptor binding (3)	2
GO:0005488	binding (1)	2
GO:0005515	protein binding (2)	2
GO:0019904	protein domain specific binding (3)	2
GO:0046914	transition metal ion binding (5)	2
GO:0016410	N-acyltransferase activity (5)	3
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen (3)	3
GO:0016740	transferase activity (2)	3
GO:0050291	sphingosine N-acyltransferase activity (6)	3

Lista dei GOID (BP)	GOterm
GO:0006355	regulation of transcription, DNA-dependent (7)
GO:0006575	cellular modified amino acid metabolic process (7)
GO:0048522	positive regulation of cellular process (4)
GO:0048523	negative regulation of cellular process (4)
GO:0009887	organ morphogenesis (3)
GO:0010033	response to organic substance (3)
GO:0030154	cell differentiation (3)
GO:0044249	cellular biosynthetic process (3)
GO:0048513	organ development (3)
GO:0048731	system development (3)
GO:0003006	developmental process involved in reproduction (2)
GO:0006950	response to stress (2)
GO:0009653	anatomical structure morphogenesis (2)
GO:0022402	cell cycle process (2)
GO:0044237	cellular metabolic process (2)
GO:0048856	anatomical structure development (2)
GO:0048869	cellular developmental process (2)
GO:0008152	metabolic process (1)
GO:0009987	cellular process (1)
GO:0032502	developmental process (1)
GO:0050896	response to stimulus (1)

Lista dei GOID (MF)	GOterm	# di DATASET
GO:0005488	binding (1)	3
GO:0005515	protein binding (2)	3
GO:0005102	receptor binding (3)	2
GO:0016740	oxidoreductase activity (2)	2
GO:0019904	oxidoreductase activity, [...] (3)	2
GO:0046914	transferase activity (2)	2
GO:0003677	DNA binding (3)	2
GO:0003824	endopeptidase activity (5)	2
GO:0016491	dioxygenase activity (3)	2
GO:0019899	nucleic acid binding (2)	2

Tabella 6.3: Lista dei termini selezionati con frequenza maggiore di 0.7 condivisi da più dataset, tra parentesi il livello del nodo. Per BP si sono riportati i nodi condivisi dai tre dataset, per MF invece la terza colonna indica il numero di dataset condivisi dal nodo.

Tabella 6.8: Tabella relativa ai cluster di geni significativi per il dataset GSE3494 analizzato per MF

GOID	GOTERM	[GSE3494] CLUSTER
GO:0005102	receptor binding (3)	1
GO:0005488	binding (1)	1
GO:0005515	protein binding (2)	1
GO:0003677	DNA binding (3)	1
GO:0019899	enzyme binding (3)	1
GO:0016491	oxidoreductase activity (2)	2
GO:0003824	catalytic activity (1)	2
GO:0004175	endopeptidase activity (5)	2
GO:0016701	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen (3)	2
GO:0016702	oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen (4)	2
GO:0016787	hydrolase activity (2)	2
GO:0016836	hydro-lyase activity (4)	2
GO:0051213	dioxygenase activity (3)	2

Tabella 6.9: Tabella relativa ai cluster di geni significativi per il dataset GSE7390 analizzato per MF

GOID	GOTERM	[GSE7390] CLUSTER
GO:0001071	nucleic acid binding transcription factor activity (1)	1
GO:0003700	sequence-specific DNA binding transcription factor activity (2)	1
GO:0003677	DNA binding (3)	2
GO:0043565	sequence-specific DNA binding (4)	2
GO:0003676	nucleic acid binding (2)	2
GO:0016491	oxidoreductase activity (2)	3
GO:0016740	transferase activity (2)	3
GO:0003824	catalytic activity (1)	3
GO:0004672	protein kinase activity (5)	3

GO:0004713	protein tyrosine kinase activity (6)	3
GO:0005083	small GTPase regulator activity (4)	4
GO:0005097	Rab GTPase activator activity (6)	4
GO:0005099	Ras GTPase activator activity (5)	4
GO:0030695	GTPase regulator activity (3)	4
GO:0005488	binding (1)	5
GO:0046914	transition metal ion binding (5)	6
GO:0005509	calcium ion binding (5)	6
GO:0046872	metal ion binding (4)	6
GO:0005515	protein binding (2)	7
GO:0008092	cytoskeletal protein binding (3)	7
GO:0019904	protein domain specific binding (3)	7
GO:0019899	enzyme binding (3)	7
GO:0042802	identical protein binding (3)	7
GO:0042803	protein homodimerization activity (4)	7
GO:0046983	protein dimerization activity (3)	7
GO:0016881	acid-amino acid ligase activity (4)	8
GO:0050662	coenzyme binding (3)	9

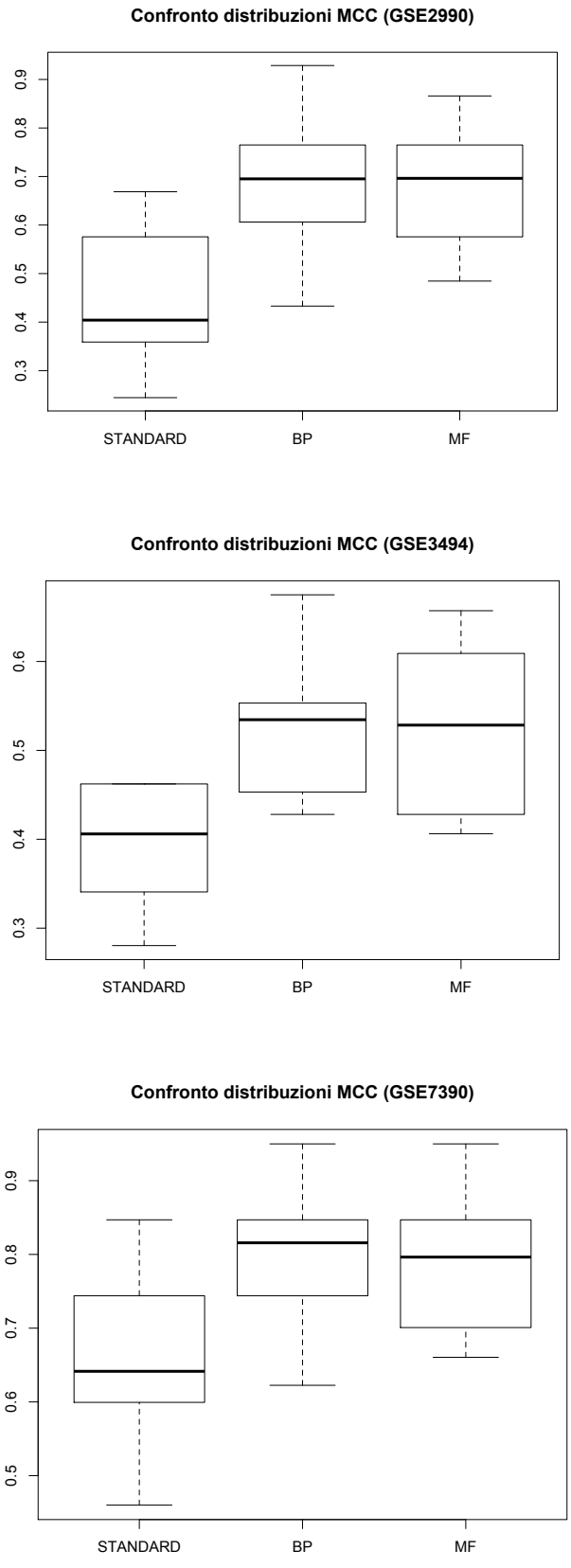

Figura 6.1: Andamento della distribuzione dei termini GO nelle varie repliche

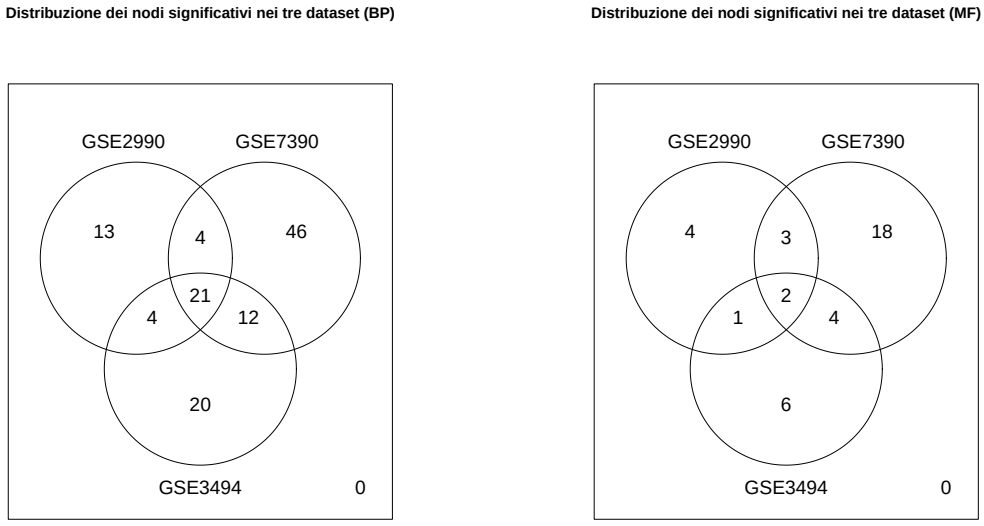

Figura 6.2: Distribuzione dei nodi significativi (frequenza maggiore di 0.7) nei tre dataset, si può notare una buona sovrapposizione tra i termini GO selezionati.

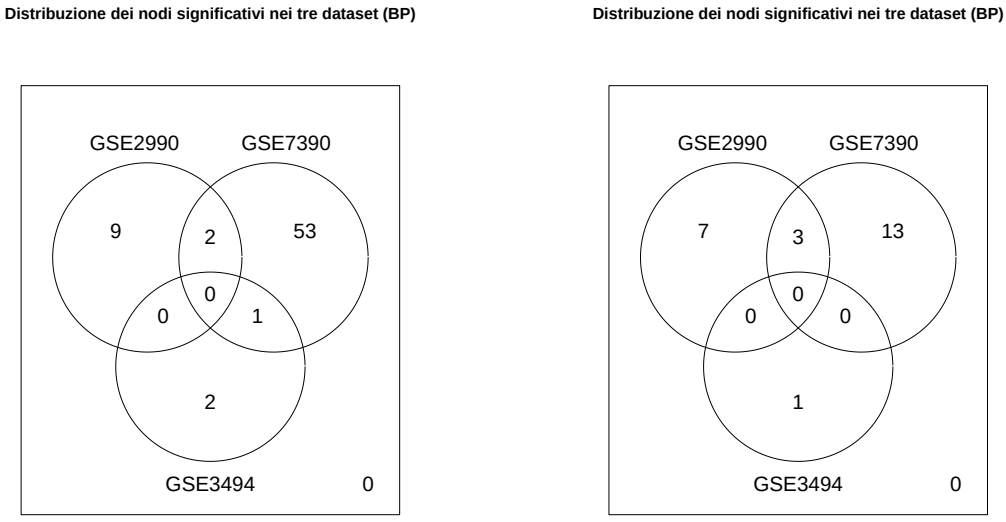

Figura 6.3: Distribuzione per il metodo standard dei nodi significativi calcolati mediante enrichment (frequenza maggiore di 0.7) nei tre dataset, si può notare come in questo caso non vi sia una buona sovrapposizione tra i termini GO selezionati e che vi siano pochi nodi selezionati, il che è indicativo della stabilità del metodo.

Capitolo 7

Conclusioni

Il presente lavoro di tesi presenta un approccio nuovo alla classificazione di dati di espressione genica, in particolare, integrando l'informazione biologica nel processo di apprendimento di un classificatore, si vuole poter trattare i geni non più come variabili indipendenti del processo di classificazione ma come espressione di un ambiente caratterizzato da forti relazioni tra le feature, come può essere l'ambiente cellulare.

Dai risultati ottenuti applicando il metodo a tre dataset differenti sullo stesso caso clinico emerge che l'idea premia anche in termini di performance di classificazione. Il metodo infatti è stato messo a confronto con un metodo di classificazione standard fornendo, in termini di MCC, risultati migliori. Questo può essere dovuto al fatto che i singoli classificatori con cui viene assemblato il classificatore finale sono costruiti considerando meno geni e quindi su un problema di dimensione minore. In questo modo si affronta anche il cosiddetto *large p, small n paradigm*. Si può dire allora che, partizionando il problema di classificazione in più sotto-problemi costruiti sulla base della GO, seppur perdendo alcuni vantaggi derivanti da un analisi multivariata, riesco a raggiungere performance di classificazione migliori.

Il contributo più alto però è relativo all'interpretabilità dei risultati. Per la sua stessa natura infatti il classificatore costruito fornisce in uscita delle liste di termini GO. In questo modo si è in grado di slegare i risultati della classificazione dalle inevitabili differenze nell'espressione genica dei singoli soggetti. Molti studi sul cancro infatti hanno dimostrato che questa patologia può essere ricondotta all'alterazione di un numero relativamente basso di processi biologici che però possono essere alterati da un altissimo numero di geni.

Diminuendo allora la risoluzione dei risultati, spostando cioè l'attenzione dai singoli geni ai processi biologici, si è in grado di avere un'immagine più nitida della patologia.

Gli sviluppi futuri del lavoro si muoveranno nel senso di incrementare la stabilità delle liste di geni in uscita, ad esempio integrando nel metodo un approccio di tipo bootstrap che permetta di rendere l'analisi più robusta e quindi i risultati più affidabili.

Un altro risultato che si cercherà di raggiungere è di associare ad ogni nodo selezionato delle liste di geni ordinate in base alla loro significatività, in questo modo il metodo sarà in grado di fornire in uscita un'informazione ben strutturata e quindi più facile da interpretare.

Statistica inferenziale

La statistica inferenziale si occupa di caratterizzare il rapporto tra una *popolazione* e un *campione* da questa estratto. Vediamo quindi nel dettaglio cosa si intende per campione.

Si consideri una popolazione finita di N unità, allora un campione casuale di ampiezza n si definisce come:

Definizione 1 *Si chiama campione casuale di ampiezza n la variabile aleatoria multipla* $(X_1, X_2, \ldots, X_n)$ *le cui componenti* $X_1, X_2, \ldots, X_n$ *associate alle varie osservazioni sono indipendenti e identicamente distribuite secondo la funzione di probabilità e di densità* $f(x)$*, essendo* $f(x)$ *il modello descrittivo della popolazione.*

Dalla definizione appena data si trae immediatamente il criterio per la determinazione della distribuzione di probabilità del campiona casuale $(X_1, X_2, \ldots, X_n)$. Infatti, dato che le $X_1, X_2, \ldots, X_n$ sono indipendenti e identicamente distribuite secondo la funzione $f(x)$, la funzione di probabilità congiunta della variabile aleatoria $(X_1, X_2, \ldots, X_n)$ è data da

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n)$$

Esempio 1 *Si prenda un campione casuale proveniente da una distribuzione normale di media* μ *e varianza* σ^2*, si avrà quindi*

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

dalla definizione di campione casuale segue allora che per un campione di n elementi si ha

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$$

Dato un campione casuale è possibile definire su questo delle funzione come ad esempio la media o la varianza, si parla in questo caso di *statistica campionaria*

Definizione 2 *Dato un campione casuale* $(X_1, X_2, \ldots, X_n)$ *si dice statistica campionaria una qualsiasi funzione del campione*

$$g(X_1, X_2, \ldots, X_n)$$

Risulta poi sotto deboli ipotesi[1] che g è una variabile aleatoria. Si noti che cadono nella definizione di statistica campionaria due funzioni molto importanti come la *media campionaria* e la *varianza campionaria* definite da

MEDIA CAMPIONARIA

$$\begin{aligned} \bar{x}_n \quad &: \quad R^n \to R \\ &\quad (X_1, X_2, \ldots, X_n) \mapsto \bar{x}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \end{aligned}$$

VARIANZA CAMPIONARIA

$$\begin{aligned} s^2 \quad &: \quad R^n \to R \\ &\quad (X_1, X_2, \ldots, X_n) \mapsto s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x}_n)^2 \end{aligned}$$

una statistica campionaria poi può essere utilizzata nella stima dei parametri di una popolazione o per la verifica di ipotesi sulla popolazione stessa a patto che se ne conosca la distribuzione di probabilià. In generale ricavare tale funzione può risultare complicato, si possono però valutare comunque alcune proprietà relativamente a media campionaria e varianza campionaria

Proposizione 1 *Data una popolazione con una distribuzione tale che*

$$E[f(x)] = \mu \quad e \quad VAR[f(x)] = \sigma^2$$

per ogni campione $(X_1, X_2, \ldots, X_n)$ *appartenente a tale popolazione risulta*

$$E[\bar{x}_n] = \mu \quad e \quad VAR[\bar{x}_n] = \frac{\sigma^2}{n}$$

Proposizione 2 ?? *Sia* $(X_1, X_2, \ldots, X_n)$ *un campiona casuale proveniente da una popolazione con distribuzione normale di media* μ *e varianza* σ^2 *segue che*

$$s^2 \sim \frac{\sigma^2}{n-1} \chi^2 (n-1)$$

dove χ^2 *indica una variabile aleatoria chi-quadro*[2]*.*

si noti per inciso che in base a questo teorema segue che

$$\frac{n-1}{\sigma^2} s^2 \sim \chi^2 (n-1)$$

[1] la funzione deve essere una funzione di Borel
[2] vedi [13]

inoltre un risultato fondamentale valido sotto opportune ipotesi

Proposizione 3 *Sia* $(X_1, X_2, \ldots, X_n)$ *un campione proveniente da una popolazione* $N(\mu, \sigma^2)$ *allora la distribuzione di probabilità della media campionaria* $\bar{x}_n$ *è una normale di media* μ *e varianza* $\frac{\sigma^2}{n}$*:*

$$\bar{x}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Le proposizione appena viste sono alla base di uno dei più importanti test di ipotesi, il *t-test*, prima di vederlo però è bene chiarire alcuni concetti fondamentali per tutti i tipi di test di ipotesi.

Bibliografia

[1] The nobel prize in physiology or medicine 1958, November 2011.

[2] L. Ding et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature, 455:1069–1075*, 2008.

[3] Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco. Elastic-net regularization in learning theory. *Computer Science and Artificial Intelligence Laboratory Technical Report*, 2008.

[4] http://www.affymetrix.com/analysis/index.affx.

[5] Siddhartha Mukherjee. *L'imperatore del male, una biografia del cancro.* Neri Pozza editore, 2011.

[6] Douglas Hanahan and Robert A. Weinberg. The hallmarks of cancer. *Cell, 100:57–70*, 2000.

[7] Hayflick Leonard. Mortality and immortality at the cellular level. *Biochemistry 62:1180–1190*, 2000.

[8] Eli Upfal, Fabio Vandin, and Benjamin J. Raphael. De novo discovery of mutated driver pathways in cancer. 2011.

[9] F. Tai and W. Pan. Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. *Bioinformatics, 23:1775-1782*, 2007.

[10] C. Lottaz and R. Spang. Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. *Bioinformatics, 21:1971-1978*, 2005.

[11] Adrian Alexa, Jorg Rahnenfuhrer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *BIOINFORMATICS, 22 no.13:1600–1607*, 2006.

[12] http://www.ncbi.nlm.nih.gov/geo/.

[13] Giuseppe Cicchitelli. *Probabilità e statistica.* Maggioli editore, 2000.

[14] The gene ontology consortium.

[15] M.H. et al. Cheok. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat. Genet., 34, 85–90*, 2003.

[16] L.J. et al. van't Veer. Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415, 530–536.*, 2002.

[17] E. et al. Huang. Gene expression predictors of breast cancer outcomes. *Lancet, 361, 1590–1596.*, 2003.

[18] E.-J. et al. Yeoh. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell, 1, 133–145*, 2002.

[19] A. et al. Bhattacharjee. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci., USA, 98, 13790–13795*, 2001.

[20] R. Tibshirani and other. Diagnosis of multiple cancer types using shrunken centroids of gene expression. *Proc. Natl Acad. Sci.*, 2002.

[21] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Statist. Soc. B 67, Part 2, pp. 301–320*, 2005.

[22] many. *Statistical Analysis of gene expression microarray data.* Terry Speed, 2003.