# VALUTAZIONE DI METODI DI GENE SELECTION PER L'ANALISI DI ESPERIMENTI CON DNA MICROARRAY

F. Ruffino (*) – G. Valentini (*) – M. Muselli (**)

(*) Dipartimento di Scienze dell'Informazione
Università di Milano
via Comelico, 39 – 20135 Milano
Tel.: +39 02 50316225
Fax: +39 02 50316373
Email: {ruffino,valentini}@dsi.unimi.it

(**) Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni
Consiglio Nazionale delle Ricerche
via De Marini, 6 – 16149 Genova
Tel.: +39 010 6475213
Fax: +39 010 6475200
Email: marco.muselli@ieiit.cnr.it
Autore di riferimento

**Sommario**

In questo lavoro viene preso in esame il problema di individuare i geni significativi (*gene selection*) per la discriminazione di due classi di tessuti attraverso l'analisi di esperimenti eseguiti con DNA microarray. In particolare, tre tecniche di gene selection molto promettenti, il metodo di Golub, le Support Vector Machines con Recursive Feature Elimination e le Switching Neural Networks con Recursive Feature Addition, sono valutate attentamente, analizzando le loro prestazioni su database artificiali ottenuti con un sistema di generazione automatico, denominato TAGGED (Technique for Automatic Generation of Gene Expression Data), capace di produrre dati biologicamente plausibili. L'impiego di database artificiali consente di pervenire ad una valutazione oggettiva dei tre metodi, non influenzata da eventuali convinzioni predeterminate. Parallelamente, le tre tecniche sono utilizzate per l'analisi di database del mondo reale; il confronto dei risultati ottenuti permette di ottenere interessanti informazioni circa i geni coinvolti nei processi fisiologici esaminati.

**Parole chiave**

DNA microarray, gene selection, machine learning, support vector machines, switching neural networks.

## 1. INTRODUZIONE

I DNA microarray sono strumenti di analisi capaci di produrre il valore del livello di espressione per decine di migliaia di geni relativi ad un determinato tessuto [1]. Il loro impiego può permetterci di giungere alla comprensione dei meccanismi che regolano un'ampia gamma di processi biologici, quali

l'insorgere di una malattia o gli effetti di un farmaco. Tuttavia, il trattamento di una così grande quantità di valori richiede l'utilizzo di strumenti opportuni di tipo statistico ed informatico, in quanto un'analisi diretta da parte di un operatore umano appare un'impresa ardua se non disperata.

A motivo della loro capacità di trattare problemi complessi di inferenza statistica multivariata, le tecniche di machine learning, come il clustering gerarchico [2], le mappe auto-organizzanti di Kohonen [3], le reti neurali [4], le support vector machines [5] e gli alberi decisionali [6], sono stati adottati con successo al fine di estrarre l'informazione contenuta in un insieme di *n* tessuti, per ognuno dei quali è stato prodotto il livello di espressione di *m* geni attraverso l'impiego di un DNA microarray. Per esempio, gli *n* tessuti differenti potrebbero essere stati prelevati da uno stesso paziente a diversi istanti di tempo, così da determinare l'effetto di un dato farmaco. Oppure, in un'altra eventualità, i tessuti si riferiscono ad *n* pazienti differenti, alcuni dei quali sono affetti da una determinata patologia.

Senza entrare nel merito dell'importante problema concernente la normalizzazione e la riduzione degli errori di acquisizione relativi ad un tipico scanner, i dati in ingresso alla tecnica di machine learning possono essere rappresentati come una matrice bidimensionale *D* a valori reali, contenente *n* righe (una per ogni tessuto) ed *m* colonne (una per ogni gene). Poiché il valore di *n* è nella migliore delle ipotesi tra 100 e 200 (a causa del costo elevato di un singolo esperimento con DNA microarray), mentre *m* è dell'ordine di qualche migliaio, la matrice *D* ha le due dimensioni molto diverse tra loro, il che influenza l'applicazione di qualunque metodo automatico di analisi.

In realtà, esistono due modi possibili per analizzare i dati contenuti nella matrice *D*: il primo di essi considera ognuna delle *m* colonne come un punto in uno spazio ad *n* dimensioni. In questo modo le coordinate di ogni punto si riferiscono ai valori di espressione che un gene specifico assume negli *n* diversi tessuti esaminati. Per esempio, quando vengono esaminati gli effetti di un farmaco, ogni coordinata rappresenta il livello di espressione del gene associato a quel punto in uno specifico istante. In questo tipo di analisi l'obiettivo è quello di trovare delle similarità tra i punti che corrispondono a geni diversi, così da stabilire dei collegamenti che aiutino a comprendere i processi biologici in atto.

Un secondo modo di analizzare la matrice *D* consiste nell'esaminare una riga alla volta, ottenendo così *n* vettori differenti che rappresentano altrettanti punti in uno spazio degli ingressi *m*-dimensionale. In questo caso le coordinate di ogni punto si riferiscono ai valori di espressione di tutti i geni associati ad uno specifico tessuto. Per esempio, se desideriamo esaminare i geni coinvolti in una determinata patologia, dobbiamo trovare le differenze tra i tessuti dei pazienti che presentano quella malattia e i tessuti di altri pazienti, considerati come controesempi. Ciò richiede la presenza di un'etichetta su ognuno degli *n* punti in esame, che identifica il corrispondente tessuto come appartenente o no ad un paziente affetto dalla patologia considerata. Il metodo automatico di analisi dovrà quindi cercare una funzione discriminante che separa in modo ottimale gli insiemi di punti appartenenti alle due classi distinte.

In questi tipi di problemi è altresì importante determinare il sottoinsieme di geni coinvolti nell'insorgere della malattia; questo permette di raggiungere due benefici immediati:

1. aumentare la conoscenza sui processi biologici in atto, in particolare i geni e i percorsi metabolici interessati,
2. migliorare la discriminazione tra i due insiemi di punti; infatti, diminuire la dimensione dello spazio degli ingressi è uno dei modi più usati per semplificare il compito di trovare la funzione discriminante ottimale.

Il problema di determinare questo sottoinsieme di geni significativi è normalmente indicato con il nome di *gene* (o *feature*) *selection*. Diversi metodi sono stati proposti in letteratura per affrontare questo tipo di problema. Golub [7] ha impiegato un semplice metodo statistico univariato ottenendo risultati interessanti nella discriminazione tra due tipi differenti di leucemia (Leukemia dataset). Più recentemente Guyon et al. [8] hanno impiegato una procedura ricorsiva, denominata Recursive Feature Elimination (RFE) e basata sull'applicazione di Support Vector Machine (SVM) lineari, per effettuare la gene selection sugli stessi dati considerati da Golub e su un ulteriore insieme di dati inerenti il problema di riconoscere la presenza del cancro al colon (Colon dataset). In entrambi i casi sono stati individuati geni significativi in accordo con la letteratura medico-biologica relativa a queste patologie.

Un'altra classe di metodi che può essere impiegata efficacemente ai fini della gene selection è quella delle *tecniche per la generazione di regole*. Esse sono capaci di risolvere un problema di classificazione, quale quello di discriminare diversi pazienti sulla base degli esiti degli esperimenti condotti con DNA microarray, generando una collezione di regole intelligibili sottostanti il processo biologico considerato. Per ispezione diretta delle regole prodotte è immediato ottenere l'insieme di geni significativi in esse impiegate. Seguendo tale approccio, gli alberi decisionali sono stati impiegati con successo per eseguire

la gene selection [6] sul Leukemia dataset e su un insieme di esperimenti relativi a diversi tipi di linfoma (Lymphoma dataset).

Alcune tecniche per la generazione di regole [9,10,11] codificano le variabili inerenti il problema di classificazione considerato in termini di stringhe binarie ed impiegano un metodo per la sintesi di funzioni booleane per eseguire la costruzione dell'insieme di regole intelligibili. Una tecnica di questo tipo è Hamming Clustering (HC)[10], il cui utilizzo nella prognosi di malattie neoplastiche [12] e nell'individuazione di segnali rilevanti in sequenze genomiche [13] ha condotto a risultati positivi. In tutte le applicazioni considerate in letteratura HC ha ottenuto valori di accuratezza paragonabili o superiori a quelli inerenti gli alberi decisionali; inoltre, le regole prodotte da HC offrono generalmente un maggior contributo informativo rispetto a quelle ottenute con gli alberi decisionali.

Per migliorare le capacità predittive di HC è stata recentemente considerata la possibilità di impiegare funzioni booleane positive al posto di quelle generiche nella costruzione dell'insieme di regole relativo ad un determinato problema di classificazione. Tale approccio risulta teoricamente perseguibile qualora venga adottato un tipo di codifica binaria che permette di eseguire opportunamente la trasformazione delle variabili continue o nominali in stringhe binarie [14]. Una tecnica specifica per la ricostruzione di funzioni booleane positive da esempi, denominata *Shadow Clustering* (*SC*) [15], consente quindi di ottenere la generazione dell'insieme di regole intelligibili e la conseguente determinazione delle variabili significative. Il dispositivo artificiale che risolve il problema di classificazione attraverso l'approccio seguito da SC viene chiamato *Switching Neural Network* (*SNN*).

Sebbene l'analisi delle proprietà teoriche e applicative delle SNN è tuttora in corso, i primi risultati ottenuti hanno mostrato che gli insiemi di regole prodotti da SC hanno generalmente un'accuratezza superiore rispetto agli analoghi insiemi prodotti da HC. In molti dei problemi analizzati le prestazioni delle SNN sono paragonabili o superiori ai migliori metodi di machine learning come le SVM o le reti neurali multistrato.

Il presente lavoro si propone pertanto di valutare la capacità di effettuare gene selection attraverso le SNN, confrontando i risultati ottenuti con quelli prodotti da altre tecniche più consolidate, quali il metodo di Golub o la RFE. Una comparazione oggettiva su dataset reali è però difficilmente realizzabile in quanto mancano le conoscenze necessarie circa l'insieme di geni effettivamente coinvolti nel processo di classificazione desiderato. Ad esempio, nel caso del Leukemia dataset la letteratura medico-biologica fornisce soltanto la conoscenza di alcuni dei geni coinvolti nei due diversi tipi di leucemia; anzi, la determinazione di nuovi geni responsabili è uno dei motivi principali dell'analisi con DNA microarray. Pertanto, se due metodi di gene selection forniscono due diversi insiemi di geni significativi, entrambi contenenti quelli già noti in letteratura, è impossibile stabilire quale delle due tecniche conduce a risultati migliori.

Per ovviare a tale problema è stata sviluppata una procedura per la generazione di dati artificiali aventi caratteristiche analoghe a quelli prodotti da DNA microarray. Tale procedura, denominata *TAGGED* (*Technique for Artificial Generation of Gene Expression Data*), è basata sul concetto di *expression signature*, introdotto nella letteratura per designare un insieme di geni correlati tra loro rispetto ad un determinato stato funzionale, dove il termine "correlati" indica che tali geni, se attivi, sono responsabili della determinazione di quello stato. Un gene presente nell' expression segnature relativa ad uno stato si definisce attivo se sovraespresso, oppure sottoespresso, rispetto ad una situazione assunta come riferimento. Pertanto, ad ogni gene può essere associata una variabile binaria, che assume il valore 1 se quel gene è attivo oppure il valore 0 in caso contrario. Impiegando tale codifica, la situazione degli $m$ geni in una cellula ad un dato istante può essere identificata da un vettore booleano $\boldsymbol{z}$ a $m$ bit, che segnala quali geni sono attivi rispetto ad uno stato funzionale di interesse.

Se associamo alla cellula considerata una variabile binaria $y$, che assume il valore 1 quando la cellula si trova nello stato funzionale considerato o il valore 0 in caso contrario, è possibile individuare, almeno in linea di principio, una funzione booleana $y = f(\boldsymbol{z})$ che lega lo stato della cellula alla situazione degli $m$ geni che la caratterizzano. I geni significativi per lo stato funzionale di interesse sono quelli che influenzano la variabile dipendente $y$ quando passano dallo stato attivo a quello inattivo o viceversa.

La procedura TAGGED utilizza un metodo biologicamente plausibile per sintetizzare due funzioni booleane $f_1(\boldsymbol{z})$ e $f_2(\boldsymbol{z})$, che si ritengono descrittive di due ipotetici stati funzionali della cellula. Successivamente viene generato un insieme di esempi $\boldsymbol{x}_j$, con $j = 1,\ldots,n$, eventualmente affetti da rumore, in modo che $n_1$ esempi corrispondano a vettori booleani $\boldsymbol{z}_j$ per i quali $f_1(\boldsymbol{z}_j) = 1$ e $n–n_1$ corrispondano a vettori $\boldsymbol{z}_j$ per i quali $f_2(\boldsymbol{z}_j) = 1$. Tali esempi simulano l'esecuzione di altrettanti esperimenti con DNA microarray su tessuti costituiti da cellule nei due stati funzionali considerati. A partire da tali esempi è possibile pensare di

valutare la qualità di un metodo di gene selection. Una misura diretta della sua bontà è data infatti dalla percentuale di geni significativi per lo stato funzionale di interesse che il metodo è in grado di individuare.

Nel prossimo paragrafo saranno descritte nel dettaglio le tecniche di gene selection considerate, evidenziando i motivi del loro impiego. Successivamente, sarà presentata la procedura TAGGED per la sintesi di dati artificiali, illustrando le diverse possibili scelte che ne caratterizzano il funzionamento. Infine, database artificiali e del mondo reale saranno impiegati per confrontare tra loro i metodi esaminati, ottenendo così una valutazione oggettiva della loro qualità.

## 2. METODI DI GENE SELECTION CONSIDERATI

Supponiamo di avere a disposizione i dati di espressione genica relativi ad $n$ tessuti, ottenuti attraverso altrettanti esperimenti con DNA microarray, ognuno dei quali produce il livello di espressione di $m$ geni preselezionati. Gli $m \cdot n$ valori reali risultanti possono essere rappresentati attraverso una matrice bidimensionale $D$, contenente $n$ righe (una per ogni tessuto) ed $m$ colonne (una per ogni gene). Ad ognuna delle $n$ righe può essere aggiunto un $(m+1)$-esimo valore binario che identifica lo stato funzionale del tessuto corrispondente (sano vs. malato, patologia 1 vs. patologia 2, ecc.).

Indicando con $\boldsymbol{x}_j$, $j = 1, \ldots, n$, il vettore reale $m$ dimensionale contenente la $j$-esima riga della matrice $D$ e con $y_j \in \{0,1\}$ il valore binario ad essa associato, si individua un insieme di $n$ coppie $(\boldsymbol{x}_j, y_j)$ che possono essere viste come il training set di un problema di classificazione binario. Risolvere tale problema consiste nel trovare una *funzione di decisione* (o *classificatore*) $f : \Re^m \to \{0,1\}$ che massimizza la probabilità di assegnare l'uscita corretta $y = f(\boldsymbol{x})$ ad un qualunque vettore $m$-dimensionale $\boldsymbol{x}$ ottenibile attraverso un esperimento con DNA microarray per l'analisi considerata.

Dato che in generale si ha $m \gg n$, il problema di classificazione sarebbe pressoché irresolubile se non si arrivasse ad individuare un sottoinsieme di geni $F \subset \{1, \ldots, m\}$ tale che la funzione di decisione $f(\boldsymbol{x})$ dipende soltanto dalle componenti $x_i$ con $i \in F$. Tale processo prende il nome di *gene* (o *feature*) *selection*; esistono svariate tecniche introdotte in letteratura per effettuare la feature selection, ma molte di esse richiedono un alto costo computazionale e non sono pertanto applicabili al trattamento di dati da DNA microarray, a causa della loro elevata dimensione.

Nel presente lavoro verrà presentato un nuovo modello connessionistico, denominato *Switching Neural Network* (*SNN*) [15], addestrabile con un metodo di classificazione di nome *Shadow Clustering* (*SC*) [16], capace di effettuare il processo di gene selection con elevata efficienza, se abbinato ad una procedura detta *Recursive Feature Addition* (*RFA*). Per valutare in modo oggettivo le sue prestazioni, i risultati ottenuti su dati artificiali e del mondo reale saranno confrontati con quelli derivanti da altre due tecniche di gene selection impiegate con successo nell'analisi di dati da DNA microarray: il metodo di Golub [7] e la procedura SVM-RFE di Guyon et al. [8].

Nei prossimi paragrafi daremo una breve descrizione delle tre tecniche considerate, rimandando agli articoli originali per un approfondimento più dettagliato.

### 2.1 Metodo di Golub

Per effettuare un'analisi dei dati da microarray inerenti due tipologie di leucemia, Golub ha proposto un metodo statistico univariato che effettua la classificazione dei tessuti a partire dai valori di espressione genica [7]. Tale metodo costruisce una graduatoria dei geni ordinandoli secondo il valore decrescente di un'opportuna misura di rilevanza $t$; il sottoinsieme $F$ sarà quindi formato prendendo i primi $g$ geni della graduatoria, dopo aver fissato in precedenza il valore di $g$. In alternativa si può scegliere un valore di soglia $q$ per la rilevanza e includere nel sottoinsieme $F$ i geni caratterizzati da un valore di rilevanza $t > q$.

Golub ha proposto di impiegare come misura della rilevanza del gene $i$-esimo la sua correlazione $t_i$ con la variabile $y$ in uscita, secondo la relazione seguente:

$$t_i = \frac{(m_i(1) - m_i(0))}{(s_i(1) + s_i(0))}$$

dove $m_i(c)$ e $s_i(c)$, per $c = 0, 1$, sono rispettivamente la media e la deviazione standard dei valori $x_i$ (per l'$i$-esimo gene) calcolate sui tessuti che appartengono alla classe $c$. Se il valore di $t_i$ è positivo, è presente una

correlazione tra l'$i$-esimo gene e la classe 1 in uscita; al contrario, se il valore di $t_i$ è negativo, è presente una correlazione tra l'$i$-esimo gene e la classe 0. Maggiore è il valore assoluto di $t_i$, maggiore è la correlazione così individuata.

Ne risulta che è possibile ottenere il sottoinsieme $F$ desiderato ordinando gli $m$ geni considerati secondo il valore decrescente di $t_i$ e prendendo i $g$ geni che si trovano ai due estremi della lista. Come sopra specificato, è possibile fissare il numero $g$ di geni o due valori di soglia $q_1 < 0$ e $q_2 > 0$ per la rilevanza, includendo in $F$ i geni aventi valore di rilevanza $t_i$ che verifica $t_i < q_1$ o $t_i > q_2$.

Si noti che il metodo calcola la rilevanza di un gene in maniera indipendentemente sia dai valori assunti dalle altre variabili di ingresso, sia dal classificatore utilizzato.

## 2.2 Metodo SVM-RFE

La formazione di una graduatoria di rilevanza per i geni può essere effettuata tenendo conto della funzione di decisione impiegata per la soluzione del problema di classificazione in esame. Un modo per realizzare questo approccio è offerto dalla procedura iterativa, detta *Recursive Feature Elimination* (*RFE*), che consta della ripetizione dei seguenti tre passi:

1. Costruzione del classificatore impiegando le sole variabili d'ingresso contenute nel sottoinsieme $F$.
2. Calcolo della rilevanza di ogni gene incluso in $F$.
3. Eliminazione dell'elemento di $F$ meno significativo.

Inizialmente si pongono in $F$ tutti gli $m$ geni che caratterizzano il problema di classificazione in esame. Successivamente, al passo 3, il sottoinsieme $F$ viene via via ridotto rimuovendo da esso il gene meno significativo. È possibile accelerare la convergenza del metodo decidendo di eliminare ad ogni iterazione un numero $l > 1$ di geni.

La procedura RFE può essere applicata quando si dispone di un metodo di classificazione capace di assegnare un valore di rilevanza ad ognuna delle variabili d'ingresso. Questo è il caso delle *Support Vector Machine* (*SVM*) lineari, impiegate con successo da Guyon et al. [8] nell'analisi di dati da DNA microarray. L'algoritmo di gene selection risultante è stato denominato *SVM-RFE*.

L'addestramento delle SVM lineari richiede che le uscite $y_j$ assumano valori nell'insieme $\{-1,+1\}$. Se indichiamo con $S^+$ (risp. $S^-$) il guscio convesso dei punti $\boldsymbol{x}_j$ con uscita +1 (risp. –1), nel caso in cui $S^+$ ed $S^-$ sono linearmente separabili è possibile costruire l'*iperpiano ottimo* $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, che ha massima distanza da tali gusci convessi. Le quantità $\boldsymbol{w}$ e $b$ si indicano rispettivamente con il nome di *vettore dei pesi* e *bias*; essi possono essere determinati attraverso la soluzione del seguente problema di programmazione quadratica convessa nelle incognite $(\boldsymbol{w},b)$:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}\cdot\boldsymbol{w}$$

$$\text{sotto il vincolo } y_j(\boldsymbol{w}\cdot\boldsymbol{x}_j+b) \geq 1, \qquad \text{per ogni } j = 1, \ldots, n.$$

Tale problema può anche essere risolto cercando i valori dei moltiplicatori di Lagrange $a_j$ nella formulazione duale di Wolfe; in questo caso abbiamo

$$\boldsymbol{w} = \sum_{j=1}^{n} a_j y_j \boldsymbol{x}_j$$

Solo i punti $\boldsymbol{x}_j$ che giacciono vicino all'iperpiano ottimo presentano un valore non nullo (positivo) del moltiplicatore di Lagrange $a_j$ corrispondente; tali punti sono detti *support vector* e catturano l'informazione essenziale circa il training set considerato. Una volta trovato l'iperpiano ottimo, chiamato anche *SVM lineare*, possiamo determinare la classe di un qualunque nuovo campione $\boldsymbol{x}$ osservando semplicemente da quale lato dell'iperpiano si trova, ovvero controllando il segno dell'espressione $\boldsymbol{w} \cdot \boldsymbol{x} + b$.

Se i due gusci convessi $S^+$ ed $S^-$ non sono linearmente separabili, la SVM lineare può ancora essere trovata, accettando che un piccolo gruppo di punti nel training set sia classificato in modo non corretto. Un fattore di regolarizzazione $C$ tiene conto del bilancio tra numero di errori e distanza dell'iperpiano dai gusci convessi dei punti correttamente classificati.

L'impiego delle SVM lineari quali classificatori per l'applicazione del metodo RFE richiede anzitutto la definizione di una misura della rilevanza $t_i$ di ciascuna componente (gene) $x_i$ del vettore $\boldsymbol{x}$ d'ingresso. Per tale scopo è possibile utilizzare il valore assoluto $|w_i|$ o il valore quadratico $w_i^2$ della corrispondente componente del vettore direzionale $\boldsymbol{w}$ nell'iperpiano ottimo; la scelta usuale è $t_i = w_i^2$.

La procedura RFE tiene pertanto in memoria un insieme $F$ contenente l'indice dei geni ritenuti significativi all'iterazione corrente. Dopo aver eseguito l'addestramento della SVM lineare, sulla base dei soli geni contenuti in $F$, rimuove da tale insieme il gene con valore di rilevanza minimo. Esso andrà ad aggiungersi ad un vettore $\boldsymbol{r}$, che conterrà pertanto gli indici dei geni con valore crescente di importanza. Inizialmente, l'insieme $F$ conterrà tutti gli indici degli $m$ geni, mentre il vettore $\boldsymbol{r}$ avrà componenti nulle.

L'intera procedura SVM-RFE può essere schematizzata secondo la seguente tabella:

Algoritmo SVM-RFE

**Input:** Training set costituito da $n$ coppie $(\boldsymbol{x}_j, y_j)$, con $j = 1, \ldots, n$.
**Output:** Vettore $\boldsymbol{r}$ contenente l'indice dei geni con valore crescente di importanza.

1. Poni $F = \{1, 2, \ldots, m\}$ e $r_i = 0$ per $i = 1, 2, \ldots, m$.
2. Sia $k = 0$. Ripeti finché $F$ non è vuoto:
   2a. Addestra una SVM lineare utilizzando i geni con indice in $F$. Sia $\boldsymbol{w}$ il vettore direzionale risultante.
   2b. Trova l'indice $i \in F$ che presenta il valore minimo della rilevanza $t_i = w_i^2$.
   2c. Incrementa il valore di $k$ e poni $r_k = i$.
   2d. Elimina l'indice $i$ dall'insieme $F$.

## 2.3 Switching Neural Networks

Le *Switching Neural Networks* (*SNN*) [15] sono semplici modelli connessionistici senza pesi, sui quali viaggiano segnali ad un solo livello; ogni unità (neurone) svolge una delle seguenti operazioni elementari: conversione A/D, prodotto logico AND, somma logica OR. È possibile dimostrare che qualunque SNN è equivalente ad un circuito digitale che realizza una funzione booleana positiva (nella quale non è necessario l'impiego dell'operatore NOT). Inoltre, le SNN sono approssimatori universali, cioè possono approssimare entro una precisione arbitraria qualunque funzione reale misurabile.

L'algoritmo di addestramento delle *SNN*, denominato *Shadow Clustering* (*SC*) [16], adotta un approccio simile a quello utilizzato dal metodo Hamming Clustering (HC) [10]; consiste dei seguenti tre passi:

1. I valori delle variabili d'ingresso (livelli di espressione genica) sono trasformati in stringhe binarie impiegando una codifica appropriata che preserva le proprietà fondamentali di ordinamento e distanza.
2. Viene ricostruita l'espressione AND-OR di una funzione booleana positiva a partire dagli esempi del training set (codificati in forma binaria).
3. La SNN corrispondente è generata direttamente a partire dall'espressione AND-OR ottenuta nel passo 2.

Poiché i valori dei livelli di espressione genica sono numeri reali, la trasformazione impiegata al passo 1 di SC richiede una preventiva discretizzazione che determina una suddivisione ottimale dei valori possibili in un numero opportuno di intervalli adiacenti. Per eseguire questa attività viene utilizzata una tecnica denominata EntMDL [17], basata su un'analisi dell'entropia degli intervalli nei quali è possibile suddividere il range di valori di ognuna delle $m$ variabili d'ingresso. Successivamente, ad ogni intervallo così individuato viene associata una stringa binaria secondo la codifica only-one, la quale è capace di preservare le proprietà di ordinamento e distanza, se una metrica appropriata è impiegata nell'insieme delle stringhe binarie.

Una volta che gli $n$ punti che formano il training set sono stati trasformati in stringhe binarie, si ottiene un insieme di coppie ingresso-uscita che può essere visto come una porzione della tavola di verità di

una funzione booleana positiva *f* incognita. Al momento non esiste in letteratura un metodo automatico per ricostruire la funzione *f* a partire da un sottoinsieme di campioni. L'algoritmo SC permette di raggiungere questo risultato cercando inoltre di generalizzare in modo soddisfacente l'informazione disponibile, così da ottenere l'uscita più probabile in corrispondenza dei punti non inclusi nel training set.

Per ottenere questo risultato SC segue un approccio competitivo, che ha lo scopo di migliorare l'accuratezza della funzione booleana positiva risultante. Ad ogni iterazione SC raggruppa stringhe binarie che appartengono alla stessa classe e sono vicine tra loro secondo la metrica considerata; in questo modo si genera un implicante primo che può essere impiegato per costruire un prodotto logico da inserire nell'espressione AND-OR risultante. Prescrizioni generali, derivanti dal contesto della teoria statistica dell'apprendimento, suggeriscono di preferire i prodotti logici che soddisfano il maggior numero di esempi nel training set e, in subordine, quelli con un minor numero di operandi. Una fase finale di pruning serve a semplificare l'espressione AND-OR finale, migliorando così la capacità di generalizzazione della SNN prodotta.

Per costruzione ognuno dei prodotti logici nell'espressione finale è associato ad un neurone nello strato nascosto della SNN e può essere tradotto immediatamente in una regola intelligibile soggiacente al problema di classificazione in esame. In questo modo, SC può essere utilizzato per produrre un insieme di regole relativo ad un determinato stato funzionale (ad es. una patologia) esaminando i livelli di espressione genica ottenuti con successivi esperimenti effettuati per mezzo di un DNA microarray.

La disponibilità di regole intelligibili che legano tra loro i geni esaminati può aiutare a comprendere i meccanismi biologici coinvolti nel problema di classificazione considerato. In aggiunta, come sottoprodotto del processo di addestramento, SC è capace di determinare eventuali variabili d'ingresso ridondanti per l'analisi in questione e di ordinare secondo un'opportuna misura di rilevanza i geni significativi. È sufficiente osservare quali geni sono utilizzati nella formazione delle regole; i geni non inclusi saranno ritenuti ridondanti, mentre quelli contenuti nelle regole più affidabili saranno associati ad un valore più elevato di rilevanza.

La presenza di più insiemi di regole equivalenti per un dato problema di classificazione suggerisce di impiegare un approccio analogo alla RFE anche per effettuare la gene selection con le SNN. Il tipo di approccio impiegato consiglia però di procedere in modo opposto alla RFE, aggiungendo al sottoinsieme *F* i geni ritenuti maggiormente importanti secondo la misura di rilevanza prescelta. La nuova metodologia verrà indicata con il nome di *Recursive Feature Addition* (*RFA*) e consta dei seguenti tre passi:

1. Costruzione del classificatore impiegando le variabili d'ingresso non incluse nel sottoinsieme *F*.
2. Calcolo della rilevanza di ogni gene non incluso in *F*.
3. Inserimento del gene più significativo in *F*.

Inizialmente l'insieme *F* sarà vuoto; ad ogni iterazione (passo 3) viene aggiunto ad *F* il gene più significativo secondo la misura di rilevanza dettata dal classificatore impiegato. Anche per RFA è possibile accelerare la convergenza decidendo di aggiungere ad ogni iterazione un numero $l > 1$ di geni; il valore di $l$ può essere fissato in precedenza, oppure si può pensare di inserire in *F* tutti i geni con rilevanza maggiore di una certa soglia $q$.

## 3. PROCEDURA TAGGED PER LA GENERAZIONE DI DATI ARTIFICIALI

L'idea di elaborare un metodo per la simulazione di dati generati da esperimenti condotti attraverso DNA-Microarray è nata dall'esigenza di conoscere a priori sia i geni responsabili dell'insorgenza di una malattia sia i gruppi di geni tra loro correlati. Infatti, se alle matrici di dati ottenuti da esperimenti reali vengono applicati metodi di gene selection o di clustering, i risultati ottenuti non possono essere valutati a meno di avere una conoscenza pregressa circa l'identità dei effettivamente coinvolti nella malattia presa in esame. Al contrario, avendo a disposizione matrici di dati dei quali si conosce completamente la struttura, un test sui metodi precedentemente citati risulta immediato.

Per questa ragione è stata elaborata la procedura *TAGGED* (*Technique for Artificial Generation of Gene Expression Data*): essa costruisce una tabella avente la stessa struttura della matrice *D* di dati ottenuta attraverso diversi esperimenti di DNA-microarray; come descritto nei paragrafi precedenti, ognuna delle $n$ righe della tabella corrisponde ad un tessuto (o ad un paziente), mentre ognuna delle $m$ colonne contiene il valore di espressione di uno stesso gene negli $n$ tessuti. Le modalità di generazione della tabella, che

chiameremo anch'essa matrice $D$ con un piccolo abuso di notazione, dipendono dal valore assegnato dall'utente ad un insieme di parametri interni al metodo.

Poiché la matrice $D$ deve riferirsi ad un problema di classificazione, dovrà contenere la simulazione di un gruppo di $n_1$ tessuti contenenti cellule in uno stato funzionale, che indicheremo con $s_1$, e di un gruppo di $n_2$ tessuti formati da cellule in un secondo stato funzionale, che indicheremo con $s_2$. Per ognuno di questi $n = n_1 + n_2$ tessuti deve essere prodotto il livello di espressione per un insieme di $m$ ipotetici geni prefissati.

La procedura TAGGED è basata sull'idea che ogni stato funzionale $s_c$, con $c = 1, 2$, è caratterizzato da un insieme di geni $F_c$ che ne costituisce l'expression signature. Per ognuno dei geni in $F_c$ esiste una condizione secondo la quale il gene risulta essere attivo ai fini dello stato funzionale $s_c$. In genere tale condizione richiede che il livello di espressione di quel gene sia maggiore (oppure minore) del livello assunto in una situazione specifica assunta come riferimento. Diremo che il gene è sovraespresso (oppure sottoespresso) nello stato funzionale $s_c$.

Pertanto, ad ognuno degli $m_c$ geni ($m_c < m$) inclusi in $F_c$ può essere associata una variabile binaria, che assume il valore 1 se quel gene è attivo oppure il valore 0 in caso contrario. Impiegando tale codifica, la situazione degli $m_c$ geni in una cellula ad un dato istante può essere identificata da un vettore booleano $\boldsymbol{z}$ a $m_c$ bit, che segnala quali geni sono attivi rispetto allo stato funzionale $s_c$ di interesse. Se associamo alla cellula considerata una variabile binaria $y$, che assume il valore 1 se la cellula si trova nello stato funzionale $s_c$ e il valore 0 in caso contrario, è possibile individuare, almeno in linea di principio, una funzione booleana $y = f_c(\boldsymbol{z})$ che lega lo stato della cellula alla situazione degli $m_c$ geni significativi che la caratterizzano.

La procedura TAGGED costruisce pertanto in maniera biologicamente plausibile la funzione $f_c(\boldsymbol{z})$ per uno stato funzionale $s_c$ considerato, producendo successivamente un insieme di $n_c$ vettori reali $\boldsymbol{x}_j$, con $j = 1, \ldots, n_c$, ad $m$ componenti, che rappresentano i valori artificiali per i livelli di espressione degli $m$ geni che caratterizzano le $n_c$ cellule simulate. Di questi $m$ geni, gli $m_c$ contenuti nel sottoinsieme $F_c$ dovranno assumere livelli di espressione che conducano a vettori binari $\boldsymbol{z}_j$ per i quali $f_c(\boldsymbol{z}_j) = 1$, mentre i restanti $m - m_c$ geni potranno assumere valori di espressione casuali all'interno del loro possibile range.

La funzione booleana $f_c(\boldsymbol{z})$ viene costruita supponendo l'esistenza di un numero $h$ di gruppi correlati di geni; la cellula sarà nello stato funzionale $s_c$ se una percentuale sufficiente $p$ degli $h$ gruppi è attiva. In modo analogo ognuno degli $h$ gruppi è formato da $l_k$ geni, con $k = 1, \ldots, h$; il $k$-esimo gruppo sarà considerato attivo se una percentuale sufficiente $q_k$ di geni al suo interno è nello stato attivo. Il sottoinsieme $F_c$ di geni significativi per lo stato funzionale $s_c$ sarà quindi costituito dagli $m_c$ geni contenuti negli $h$ gruppi impiegati per la costruzione della funzione $f_c(\boldsymbol{z})$. Per ognuno di essi devono essere stabiliti due parametri $u_i$ e $v_i$; l'$i$-esimo gene sarà nello stato attivo se il suo livello di espressione $x_i$ soddisfa la disuguaglianza $x_i > u_i$, nel caso $v_i > 0$, o la disuguaglianza $x_i < -u_i$, nel caso $v_i < 0$.

Per chiarire il processo di costruzione di un insieme di $n_c$ esempi per uno stato funzionale $s_c$, supponiamo che la funzione booleana $f_c(\boldsymbol{z})$ sia costruita partendo da $h = 4$ gruppi di geni, indicati con $G_k$, $k = 1, 2, 3, 4$. Se il numero totale di geni è $m = 20$, poniamo che i gruppi $G_k$ siano così formati:

$$G_1 = \{10, 12, 14, 15\}\ , \quad G_2 = \{9, 14, 20\}\ , \quad G_3 = \{3, 8, 15\}\ , \quad G_4 = \{3, 10, 13, 19\}$$

Come si può notare due diversi gruppi possono contenere lo stesso gene. Il sottoinsieme $F_c$ avrà pertanto la forma seguente: $F_c = \{3, 8, 9, 10, 12, 13, 14, 15, 19, 20\}$, con $m_c = 10$.

Supponiamo ora che le percentuali $q_k$ siano tutte pari a 0.3 (cioè almeno il 30% dei geni di un gruppo deve essere nello stato attivo, perché quel gruppo sia definito attivo), mentre $p = 0.5$ (almeno metà dei gruppi deve essere attivo perché la cellula sia nello stato funzionale $s_c$). Inoltre, assegniamo ai parametri $u_i$ e $v_i$, relativi ad ognuno dei geni in $F_c$ i valori:

$$u_3 = 0.25\ , \quad u_8 = 0.12\ , \quad u_9 = -0.56\ , \quad u_{10} = 0.65\ , \quad u_{12} = -0.17,$$
$$u_{13} = -0.36\ , \quad u_{14} = 0.42\ , \quad u_{15} = -0.21\ , \quad u_{19} = -0.58\ , \quad u_{20} = 0.46$$
$$v_3 = v_8 = v_{10} = v_{14} = v_{20} = 1\ , \qquad v_9 = v_{12} = v_{13} = v_{15} = v_{19} = -1$$

In base a tali parametri il gene n. 8 sarà attivo se $x_8 > 0.12$; analogamente il gene n. 15 sarà attivo se $x_{15} < -0.21$.

Se per ognuno dei 20 geni il livello di espressione $x_i$ può assumere valori nell'intervallo $[-1,+1]$, una cellula caratterizzata dai seguenti valori:

$x_1 = 0.76$ , $x_2 = -0.11$ , $x_3 = -0.32$ , $x_4 = -0.58$ , $x_5 = 0.92$ , $x_6 = 0.08$ , $x_7 = -0.27$ ,
$x_8 = 0.02$ , $x_9 = -0.16$ , $x_{10} = 0.81$ , $x_{11} = 0.43$ , $x_{12} = 0.79$ , $x_{13} = 0.61$ , $x_{14} = 0.52$ ,
$x_{15} = -0.10$ , $x_{16} = -0.34$ , $x_{17} = 0.75$ , $x_{18} = -0.20$ , $x_{19} = -0.69$ , $x_{20} = -0.36$

sarà nello stato funzionale $s_c$. Infatti, poiché $x_{10} > 0.65$, $x_{14} > 0.42$ e $x_{19} < -0.58$, i gruppi $G_1$ e $G_4$ sono attivi, cioè un numero pari alla metà di $h$.

Poiché il problema di classificazione necessita di un training set con cellule appartenenti a due stati funzionali la procedura TAGGED produce due funzioni $f_1$ ed $f_2$, relative a due diversi stati funzionali $s_1$ ed $s_2$; per ognuna di esse genera un numero $n_c$ di vettori $m$-dimensionali $\boldsymbol{x}_j$ contenenti i valori artificiali dei livelli di espressione genica. Il grado di flessibilità è elevato: infatti, è possibile selezionare in modo indipendente i valori da assegnare alle variabili $n$, $m$, $n_1$, $n_2$, $h$, $p$, $q_k$, nonché esprimere degli intervalli per la dimensione dei gruppi $G_k$ e per i parametri $u_i$, $v_i$, all'interno del quale sarà scelto in modo casuale il valore effettivo da impiegare.

È inoltre possibile specificare se due gruppi diversi possono contenere geni comuni e permettere l'esistenza di gruppi $G_k$ necessari, la cui attivazione è indispensabile perché la cellula assuma lo stato funzionale considerato. Infine, si può controllare il grado di somiglianza delle due funzioni $f_1$ ed $f_2$, impostando il numero di gruppi di geni comuni (necessari o no) presenti in esse. L'errore di acquisizione insito negli esperimenti con DNA microarray può essere simulato agendo su un parametro $E$ che rappresenta la probabilità con cui una cellula in uno dei due stati funzionali considerati può essere erroneamente assegnata alla classe opposta.

## 4. RISULTATI OTTENUTI NELL'ANALISI DI DATI ARTIFICIALI E DEL MONDO REALE

La parte sperimentale è stata sviluppata confrontando i risultati dei tre metodi di gene selection, metodo di Golub (GOLUB), SVM-RFE e SNN-RFA, su tre diversi dataset reali, contenenti valori di espressione genica ottenuti attraverso esperimenti eseguiti con DNA microarray su diversi tessuti:

- ✓ Leukemia [7]: prende in esame il problema della distinzione tra due varianti di leucemia, AML (Acute Myeloid Leukemia) e ALL (Acute Lymphoblastic Leukemia). I dati sono relativi a 72 tessuti suddivisi in 47 casi di ALL e 25 casi di AML; ogni esperimento analizza il livello di espressione di 7129 geni per un dato paziente. La matrice risultante, costituita da 72 righe e 7129 colonne, è stata quindi partizionata in una matrice di training di 38 tessuti e una matrice di test di 34 tessuti, secondo quanto riportato in [7].

- ✓ Colon cancer [18]: costituito da un totale di 62 tessuti, di cui 22 sani e 40 colpiti da cancro al colon, per ognuno dei quali sono stati analizzati 2000 geni. La matrice di training è stata creata in modo casuale estraendo 31 tessuti, così da mantenere le proporzioni originali (11 tessuti sani e 20 malati).

- ✓ Lymphoma [19]: contiene l'analisi di 4026 geni su 96 tessuti, di cui 46 colpiti da Diffuse Large B-Cell Lymphoma (DLBCL), 11 da B-cell Chronic Lymphocytic Leukemia (B-CLL), 9 da Follicular Lymphoma (FL) e i restanti 24 normali. L'obiettivo in questo caso è discriminare DLBCL dalle altre due patologie e dai tessuti normali. Procedendo come per il dataset Colon, dalla matrice originale è stata estratta in modo casuale una matrice di training contenente 48 righe e 4026 colonne.

Parallelamente alla ricerca dei geni significativi relativi ai casi appena descritti, i tre metodi di gene selection sono stati applicati, allo scopo di valutarne le prestazioni, a tre matrici artificiali, create per mezzo della procedura TAGGED. Infatti, come descritto nei precedenti paragrafi, la procedura TAGGED permette di costruire matrici di dati che simulano dataset ottenuti con la tecnica dei DNA microarray, in cui i geni coinvolti nella determinazione dei due stati funzionali sono completamente noti.

Le tre matrici artificiali sono state costruite in modo da avere lo stesso numero di righe e colonne delle tre matrici reali, facendo sì che le due classi contengano una quantità di esempi uguale a quella del caso

reale corrispondente. Inoltre le matrici artificiali sono state partizionate in una matrice di training e una di test, con le stesso criterio utilizzato per la suddivisione delle matrici reali. Abbiamo quindi che la matrice del dataset Leukemia artificiale è costituita da 72 righe e 7129 colonne, in cui 47 righe appartengono alla classe 1 mentre le restanti 25 appartengono alla classe 0. Dalla matrice globale si ottiene in modo casuale la matrice del training set contenente 38 righe. In modo analogo sono stati costruiti i dataset Colon artificiale e Lymphoma artificiale.

Oltre a queste somiglianze, poco significative per poter affermare che i dataset artificiali e i corrispondenti dataset reali siano effettivamente simili, i dati artificiali sono stati creati in modo che le curve relative all'accuratezza ottenuta da due classificatori (la SVM-Lineare e il metodo di Golub[7]), al variare del numero dei geni considerati, abbiano lo stesso andamento sia nei dataset reali che in quelli artificiali corrispondenti.

Per ottenere dati artificiali con le caratteristiche richieste, la procedura TAGGED ha creato le tre matrici utilizzando funzioni booleane $f_1$ ed $f_2$ strutturate come segue:

- ✓ Dataset Leukemia artificiale:

  La funzione $f_1$ è basata su 7 gruppi. La percentuale di attivazione di ogni gruppo è stata fissata al 70% mentre, affinché $f_1$ assuma valore 1, occorre che tutti i gruppi siano attivi. La seconda funzione è composta da 8 gruppi. Anche in questo caso occorre che tutti i gruppi siano attivi perché $f_2$ assuma valore 1, mentre è sufficiente che l'80% dei geni all'interno di un gruppo sia attivo perché tale gruppo sia attivo. Le due funzioni hanno inoltre 4 gruppi comuni. Il numero totale dei geni presenti nei gruppi che costituiscono le due funzioni è 121.

- ✓ Dataset Colon artificiale:

  Le funzioni $f_1$ ed $f_2$ sono composte rispettivamente da 4 e da 5 gruppi che devono essere tutti attivi perché la corrispondente funzione assuma valore 1. Affinché un gruppo di $f_1$ sia attivo occorre che il 60% dei suoi geni sia attivo, mentre per $f_2$ è necessario che siano attivi il 70% dei geni. Le due funzioni non hanno alcun gruppo comune e il totale dei geni appartenenti ai gruppi è 218.

- ✓ Dataset Lymphoma artificiale:

  La funzione $f_1$ è basata su 8 gruppi, nessuno dei quali è necessario per la sua attivazione. Affinché $f_1$ assuma valore 1 occorre che almeno 7 gruppi siano attivi e perché un gruppo sia attivo è necessario che più del 70% dei geni che lo compongono sia attivo. La funzione $f_2$ è costituita da 9 gruppi e la percentuale necessaria di gruppi attivi per ottenere $f_2(\mathbf{z}) = 1$ è del 70%, mentre per attivare i singoli gruppi la percentuale di geni attivi deve superare il 90%. In questo caso non esistono gruppi comuni tra le due funzioni e il numero totale di geni coinvolti è 439.

I tre metodi di gene selection qui considerati sono stati confrontati su ognuna delle sei matrici complete (tre reali e tre artificiali) e sui sei rispettivi training set ottenuti nel modo sopra descritto. In particolare per ogni dataset ogni metodo ha estratto i 200 geni più rilevanti ed è stata quindi valutata la percentuale di geni comuni ottenuta intersecando a due a due gli insiemi elaborati dai tre metodi, applicati sia all'intero dataset che al corrispondente training set. In questo modo è stato possibile un confronto tra i geni rilevati dallo stesso metodo su matrici di dimensioni diverse. Inoltre, poiché nei tre casi artificiali i geni realmente coinvolti nelle funzioni $f_1$ ed $f_2$ sono completamente noti, si sono messi in relazione i primi 200 geni selezionati dai tre metodi (sia utilizzando l'intera matrice che soltanto il training set) con i geni veramente significativi per la determinazione dei due stati.

Le percentuali dei geni comuni sono state raccolte in tre tabelle, la prima relativa ai dataset Leukemia reale e Leukemia artificiale (Tab. 1), la seconda per Colon reale e Colon artificiale (Tab. 2) e l'ultima relativa a Lymphoma reale e Lymphoma artificiale (Tab. 3). I valori, come già anticipato, rappresentano le percentuali di sovrapposizione tra i primi 200 geni delle graduatorie elaborate dai tre diversi metodi di gene selection sia sull'intera matrice dei dati (GOLUB Totale, SVM-RFE Totale e SNN-RFA Totale) che sul solo training set (GOLUB Training, SVM-RFE Training e SNN-RFA Training).

Tab. 1: Percentuale di geni comuni ottenuti con i tre metodi di gene selection considerati sui dataset Leukemia reale ed artificiale.

Tab. 1: Percentuale di geni comuni ottenuti con i tre metodi di gene selection considerati sui dataset Leukemia reale ed artificiale.

| | | | GOLUB Training | SNN-RFA Training | SVM-RFE Totale | GOLUB Totale | SNN-RFA Totale | ***LEUKEMIA REALE*** |
|---|---|---|---|---|---|---|---|---|
| | | | 39.5% | 28.5% | 35% | 30% | 25.5% | SVM-RFE Training |
| GOLUB Training | 37% | | | **44%** | 26% | **54.5%** | 36.5% | GOLUB Training |
| SNN-RFA Training | 24% | **44%** | | | 23% | 37% | 33.5% | SNN-RFA Training |
| SVM-RFE Totale | 33.5% | 23.5% | 15.5% | | | 41.5% | 40% | SVM-RFE Totale |
| GOLUB Totale | 28% | **50.5%** | 35.5% | 29% | | | **59.5%** | GOLUB Totale |
| SNN-RFA Totale | 22.5% | 41% | 34.5% | 19% | **49%** | | | |
| Geni realmente significativi | 23% | **37%** | 32% | 18.5% | **40.5%** | 39% | | |
| ***LEUKEMIA ARTIFICIALE*** | SVM-RFE Training | GOLUB Training | SNN-RFA Training | SVM-RFE Totale | GOLUB Totale | SNN-RFA Totale | | |

Nella porzione di Tab. 1 relativa al dataset Leukemia artificiale risalta la significativa sovrapposizione (50.5%) tra i geni selezionati dal GOLUB sull'intera matrice artificiale e i geni selezionati dallo stesso metodo utilizzando soltanto il training set. Come si può notare, un analogo comportamento caratterizza i risultati relativi al dataset Leukemia reale, dove tale sovrapposizione raggiunge il valore di 54.5%. Il migliore accordo tra due diversi metodi si ottiene, nel caso di analisi sull'intero dataset, tra il vettore ottenuto attraverso GOLUB e quello elaborato dal metodo SNN-RFA, sia per quanto riguarda il dataset reale che quello artificiale. Situazione analoga si verifica se si confrontano i risultati ottenuti da due diversi metodi nel caso di analisi sul solo training set.

Relativamente alla matrice artificiale, per la quale sono noti i geni significativi, è interessante notare le buone percentuali di accordo tra i diversi metodi e i geni effettivamente coinvolti nelle funzioni che determinano i due stati. In particolare, GOLUB ed SNN-RFA ottengono i migliori risultati contando tra i primi 200 geni, selezionati utilizzando l'intera matrice artificiale, rispettivamente 81 (40.5%) e 78 (39%) elementi dei gruppi relativi alle due funzioni. SVM-RFE risulta invece notevolmente meno efficace, arrivando a trovare soltanto 37 (18.5%) geni significativi. Analogo andamento si riscontra nell'analisi effettuata con la sola matrice di training, dove GOLUB e SNN-RFA risultano essere ancora i metodi migliori, sebbene il distacco da SVM-RFE diminuisca sensibilmente.

Anche nella Tab. 2, relativa ai risultati ottenuti applicando i tre metodi di gene selection considerati ai dataset Colon reale e artificiale, si può notare un buon accordo tra i valori pertinenti le analisi di gene selection sulla matrice reale e su quella generata artificialmente. Infatti, sia nel caso reale che nel caso artificiale la percentuale di geni comuni è massima quando si considera l'intersezione tra i risultati di GOLUB e quelli di SNN-RFA, sia per quanto riguarda l'analisi sull'intera matrice (85% per la matrice artificiale e 73.5% su quella reale), sia per quanto riguarda l'analisi sulla sola matrice di training (81.5% per la matrice reale e 66% per la matrice artificiale). Se invece confrontiamo le percentuali di geni comuni ottenute dallo stesso metodo applicato alla matrice totale e all'insieme di training, GOLUB ed SNN-RFA risultano essere ancora una volta le tecniche più stabili.

Osservando infine la sovrapposizione tra le collezioni di geni selezionati dai tre metodi nel caso del dataset artificiale e l'insieme dei geni realmente significativi, troviamo che GOLUB ed SNN-RFA forniscono risultati notevolmente migliori del metodo SVM-RFE. In particolare, quando viene considerata l'intera matrice di dati SNN-RFA ottiene la percentuale migliore, raggiungendo il 93.5%.

Anche nella Tab. 3, che contiene le percentuali di geni comuni selezionati dai tre metodi di gene selection relativamente ai dataset Lymphoma reale e artificiale, è possibile rilevare forti affinità tra i risultati ottenuti con la matrice reale e con quella artificiale. Come nelle due tabelle precedenti la maggiore

sovrapposizione si riscontra con GOLUB ed SNN-RFA, sia nel caso di analisi sull'intera matrice che sulla matrice di training. Essi ritrovano anche in questo caso il più alto numero di geni realmente significativi per il dataset artificiale, raggiungendo la percentuale massima del 100% quando viene analizzata la matrice complessiva dei dati.

Tab. 2: Percentuale di geni comuni ottenuti con i tre metodi di gene selection considerati sui dataset Colon reale ed artificiale.

| | | | GOLUB Training | SNN-RFA Training | SVM-RFE Totale | GOLUB Totale | SNN-RFA Totale | ***COLON REALE*** |
|---|---|---|---|---|---|---|---|---|
| | | | 41% | 35.5% | 36.5% | 31% | 32.5% | SVM-RFE Training |
| GOLUB Training | 28.5% | | | **66%** | 22.5% | **66.5%** | 62.5% | GOLUB Training |
| SNN-RFA Training | 27% | **81.5%** | | | 18% | 52.5% | 58% | SNN-RFA Training |
| SVM-RFE Totale | 59% | 28.5% | 23% | | | 25% | 24.5% | SVM-RFE Totale |
| GOLUB Totale | 26% | **81%** | 78.5% | 26% | | | **73.5%** | GOLUB Totale |
| SNN-RFA Totale | 25.5% | 78.5% | 78% | 24% | **85%** | | | |
| Geni realmente significativi | 32.5% | **80.5%** | 78% | 30% | 91.5% | **93.5%** | | |
| ***COLON ARTIFICIALE*** | SVM-RFE Training | GOLUB Training | SNN-RFA Training | SVM-RFE Totale | GOLUB Totale | SNN-RFA Totale | | |

Tab. 3: Percentuale di geni comuni ottenuti con i tre metodi di gene selection considerati sui dataset Lymphoma reale ed artificiale.

| | | | GOLUB Training | SNN-RFA Training | SVM RFE Totale | GOLUB Totale | SNN-RFA Totale | ***LYMPHOMA REALE*** |
|---|---|---|---|---|---|---|---|---|
| | | | 21% | 17% | 54.5% | 18.5% | 16.5% | SVM RFE Training |
| GOLUB Training | 26% | | | **40%** | 17.5% | **63.5%** | 26.5% | GOLUB Training |
| SNN-RFA Training | 17.5% | **56.5%** | | | 12.5% | 31% | 32% | SNN-RFA Training |
| SVM RFE Totale | 47.5% | 15.5% | 24.5% | | | 17.5% | 14.5% | SVM RFE Totale |
| GOLUB Totale | 17.5% | **91%** | 57% | 17.5% | | | **50.5%** | GOLUB Totale |
| SNN-RFA Totale | 19% | 73% | 65% | 21% | **76.5%** | | | |
| Geni realmente significativi | 32% | **100%** | 95.5% | 31% | **100%** | **100%** | | |
| ***LYMPHOMA ARTIFICIALE*** | SVM RFE Training | GOLUB Training | SNN-RFA Training | SVM RFE Totale | GOLUB Totale | SNN-RFA Totale | | |

## 5. CONCLUSIONI

Le analisi effettuate mostrano l'importanza della tecnica TAGGED, ai fini della valutazione oggettiva dei metodi di gene selection. Confrontando i risultati ottenuti su dataset reali e sui loro corrispettivi artificiali si notano forti analogie, riscontrabili nelle tabelle presentate al precedente paragrafo. Questo attesta che la procedura TAGGED, basandosi sull'idea di Expression Signature, ha la capacità di fornire dataset artificiali con caratteristiche analoghe, dal punto di vista statistico, a quelle dei dataset reali. Questa interessante proprietà consente di valutare in modo oggettivo i metodi di gene selection, indipendentemente dal procedimento seguito per determinare il sottoinsieme di geni significativi.

Una tale operazione risulta impossibile nel caso in cui ci si limiti ad analizzare dataset reali, in quanto non si conoscono in generale i geni che caratterizzano un determinato stato funzionale per una cellula. In particolare, i risultati presentati al paragrafo precedente, ottenuti applicando tre diversi metodi di gene selection, GOLUB, SVM-RFE e SNN-RFA, a tre dataset artificiali generati in accordo con altrettanti dataset del mondo reale, hanno mostrato la superiorità di GOLUB ed SNN-RFA nel determinare i geni significativi per discriminare tra classi di cellule con diverso stato funzionale. Analizzando soltanto alcune decine di campioni artificiali prodotti da TAGGED in relazione a funzioni binarie dipendenti da più di un centinaio di variabili (geni), GOLUB ed SNN-RFA sono stati capaci di determinare una percentuale rilevante di ingressi significativi (dal 40.5% nel caso del dataset Leukemia artificiale al 100% del dataset Lymphoma artificiale).

È interessante notare che i geni ottenuti da entrambi i metodi sono significativi con elevata probabilità; infatti nel dataset Leukemia artificiale il 78.6% dei geni comuni è realmente significativo. Le analoghe percentuali per Colon artificiale e Lymphoma artificiale sono significativamente superiori, attestandosi al 99.41% e al 100%, rispettivamente. Bisogna rilevare, a tale proposito che GOLUB, essendo un metodo statistico univariato non è in grado di cogliere le correlazioni tra i diversi geni, a differenza di SNN. Questa limitazione ha ricadute dirette nella diversa accuratezza dei due metodi in fase di classificazione; infatti SNN raggiunge generalmente valori di accuratezza migliori rispetto a GOLUB.

Sorprendono invece i risultati mediocri ottenuti da SVM-RFE nell'analisi effettuata, anche in considerazione del fatto che le sue prestazioni in fase di classificazione sono migliori di quelle relative a GOLUB [8]. Questa apparente incongruenza può essere spiegata notando che SVM-RFE ottiene i migliori valori di accuratezza quando utilizza un limitato sottoinsieme di geni; probabilmente tale sottoinsieme include soltanto alcuni geni significativi per il problema di classificazione in esame.

## BIBLIOGRAFIA


[1] M. L. T. LEE: "Analysis of microarray gene expression data", Kluwer Academic Publishers, 2004.

[2] M. B. EISEN ET AL.: "Cluster analysis and display of genome-wide expression patterns", Proceedings of the National Academy of Science USA, **95**, 14863–14868, 1998.

[3] T. KATO ET AL.: "Analysis of DNA microarray data by using self-organizing maps", Genome Informatics, **14**, 328–329, 2003.

[4] J. KHAN ET AL.: "Classification and diagnostic prediction of cancer using gene expression profiling and artificial neural networks", Nature Medicine, **7**, 673–679, 2001.

[5] M. P. S. BROWN ET AL.: "Knowledge-based analysis of microarray gene expression data using Support Vector Machines", Proceedings of the National Academy of Science USA, **97**, 262–267, 2000.

[6] K. FACELI ET AL. "Evaluation of gene selection metrics for tumor cell classification", Genetics and Molecular Biology, **27**, 651–657, 2004.

[7] T. R. GOLUB ET AL.: "Monotone molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, **286**, 531–537, 1999.

[8] I. GUYON ET AL.: "Gene selection for cancer classification using support vector machines", Machine learning, **46,** 389–422, 2002.

[9] E. BOROS ET AL.: "An implementation of Logical Analysis of Data", IEEE Transactions on Knowledge and Data Engineering, **12,** 292–306, 2000.

[10] M. MUSELLI, D. LIBERATI: "Binary rule generation via Hamming Clustering", IEEE Transactions on Knowledge and Data Engineering, **14,** 1258–1268, 2002.

[11] S. J. HONG: "R-MINI: An alternative approach for generating minimal rules from examples", IEEE Transactions on Knowledge and Data Engineering, **9**, 709–717, 1997.

[12] G. PAOLI, M. MUSELLI, R. BELLAZZI, R. CORVÓ, D. LIBERATI, F. FOPPIANO: "Hamming Clustering techniques for the identification of prognostic indices in patients with advanced head and neck cancer treated with radiation therapy", Medical & Biological Engineering & Computing, **38**, 483–486, 2000.

[13] P. FERRO, A. FORLANI, M. MUSELLI, U. PFEFFER: "Alternative splicing of the human estrogen receptor a primary transcript: Mechanisms of exon skipping", International Journal of Molecular Medicine, **12,** 355–363, 2003.

[14] M. MUSELLI: "Monotone Boolean functions are universal approximators", Rapporto interno IEIIT/GE/1/03, C.N.R. - Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, 2003.

[15] M. MUSELLI: "Switching Neural Networks: A new connectionist model for classification", Rapporto interno IEIIT/GE/1/05, C.N.R. - Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, 2005.

[16] M. MUSELLI, A. QUARATI: "Shadow Clustering: A method for monotone Boolean function synthesis", Rapporto interno IEIIT/GE/2/03, C.N.R. - Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, 2003.

[17] R. KOHAVI, M. SAHAMI: "Error-based and entropy-based discretization of continuous features", International Conference on Machine Learning, 194–202, Tahoe City, CA, 1995.

[18] U. ALON ET AL.: "Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proceedings of the National Academy of Science USA, **96,** 6745–6750, 1999.

[19] A. A. ALIZADEH ET AL.: "Different types of diffuse large B-cell lymphoma identified by gene expression profiling", Nature, **403,** 503–511, 2000.