Mathematica Italia
User Group Meeting

Uso di *Mathematica* per la classificazione di dati di qualità variabile

Dario Malchiodi, Matteo Re e Giorgio Valentini

Università degli Studi di Milano

Gli algoritmi di apprendimento supervisionato hanno il fine di approssimare una funzione sulla base dell'osservazione di un insieme di esempi, intesi come coppie (argomento, valore) solitamente affette da errore di misurazione. Gli approcci alla base di questo tipo di algoritmi assumono tipicamente che i dati di partenza siano omogenei. Nell'ambito della classificazione sulla base di esempi, questo lavoro descrive un approccio alternativo che, al contrario, presuppone una diversa qualità per ogni dato a disposizione, così che sia possibile stabilire un ranking degli esempi a partire dalla loro importanza.

Sulla base di una valutazione quantitativa di questa importanza viene descritta una famiglia di algoritmi di classificazione basati su vettori di supporto. In questi algoritmi l'informazione di partenza include la sopra menzionata valutazione della qualità dei singoli esempi, al fine di promuovere la corretta descrizione degli esempi più importanti a scapito di quelli rimanenti.

L'approccio viene implementato accoppiando Wolfram *Mathematica* con un ottimizzatore numerico general-purpose, al fine di applicarlo sia a dati sintetici che per risolvere problemi di riconoscimento di cifre manoscritte sulla base della loro qualità e di classificazione di proteine sulla base delle loro annotazioni.

Introduzione

Il problema della classificazione [21] consiste nell'apprendere come assegnare a due o più *classi* prefissate degli *oggetti* appartenenti a un dato insieme. L'apprendimento viene fatto sulla base di un insieme di *esempi*, ognuno costituito da un oggetto x e da un'*etichetta* y che individua la classe a cui x deve essere assegnato. L'output della procedura è un *classificatore*, inteso come funzione dall'insieme degli oggetti a quello delle etichette, che permette quindi di ottenere l'etichetta, e dunque la classe, di un generico oggetto (non necessariamente utilizzato al fine di determinare il classificatore). Questo procedimento — se vogliamo, meccanico — di apprendimento affonda quindi le sue radici su di un'analogia con i processi di apprendimento alla base del ragionamento umano. A differenza di questi ultimi, però, considera uniformemente gli esempi a disposizione, dando loro la medesima importanza, quando invece il ragionamento umano può gestire tutta una gamma di incertezze negli esempi a disposizione [14], per esempio a causa di una diversa attendibilità delle fonti da cui gli esempi provengono [18], oppure in funzione di una diversa collocazione temporale nell'acquisizione dell'informazione. Anche nei campi applicativi in cui il problema della classificazione viene tipicamente applicato e risolto esistono svariati esempi di situazioni in cui i dati a disposizione non hanno necessariamente la stessa importanza o la stessa qualità. Per esempio, nell'ambito del riconoscimento di volti vi sono immagini inerentemente più difficili da trattare rispetto ad altre: si pensi alla presenza di occhiali o di barba e baffi [15]. Analogamente, molti dati reperibili su Internet nell'ambito delle reti sociali sono automaticamente dotati di misure di qualità sotto forma di *rating* [16] assegnati dagli utenti stessi (la soddisfazione media dei compratori che hanno acquistato oggetti da un dato venditore nei siti di *e-auction*, oppure la valutazione dell'affidabilità data dagli affittuari di appartamenti o di stanze in albergo e così via). Più in generale, vi sono situazioni in cui la raccolta dei dati non può giustificare la tipica assunzione di indipendenza e descrivibilità tramite una stessa distribuzione di probabilità, come nei casi caratterizzati da una differente varianza degli errori di misurazione (dati eteroschedastici [3]) o dalla presenza di valori che veicolano informazione di carattere parziale (dati censurati [13]).

Il concetto di qualità associata ai dati è universalmente riconosciuto come una nozione con varie sfaccettature [22, 24], per le quali tra l'altro non si ha nella letteratura esistente una descrizione piena e condivisa. Analizzando i principali lavori che trattano di qualità dei dati è facile arrivare a contare oltre venti dimensioni su cui misurare questo concetto, alcune delle quali legate ad aspetti particolari del ciclo di vita dell'informazione: si consideri per esempio la concisione, l'ambito, o anche la possibilità di effettuare confronti valutando la disponibilità di particolari licenze o di opportuni formati per i dati. In ogni caso, le più importanti osservabili legate alla qualità dei dati sono universalmente riconosciute come quelle originariamente elencate da Feltham [5]:

- la *rilevanza*, intesa come capacità dei dati osservati di descrivere pienamente i fenomeni che vengono studiati, misurata in termini del grado in cui l'informazione veicolata è effettivamente collegata ai problemi che si vogliono risolvere;
- l'*accuratezza*, vista come quantificazione della differenza tra il valore effettivo e quello osservato, vuoi legata al processo di misurazione, vuoi a causa di dati ottenuti da diverse fonti, o ancora espressi in termini di proprietà soggettive valutate da differenti individui;
- la *tempestività* (anche indicata come *puntualità*), identificata con la misura del lasso di tempo intercorso tra il processo di misurazione e l'utilizzazione del valore misurato, da valutare in contesti caratterizzati da dinamicità di informazione.

In questo lavoro adotteremo l'ipotesi semplificativa che ogni dato a disposizione sia accoppiato con un valore numerico che riassuma la qualità del dato stesso, e analizzeremo come un noto algoritmo di classificazione possa venire esteso sfruttando nella sua formulazione i valori che descrivono la qualità dei dati.

Il lavoro è organizzato come segue: il prossimo paragrafo descrive l'algoritmo di classificazione preso in considerazione, mentre quello che segue indica come tale algoritmo può essere modificato in modo da sfruttare la qualità dei dati. Segue un paragrafo dedicato alla descrizione dei risultati ottenuti in vari ambiti sperimentali. Il lavoro termina con alcune brevi considerazioni conclusive.

Classificazione tramite vettori di supporto

L'approccio alla classificazione basato su vettori di supporto [4] richiede che gli oggetti da classificare siano descrivibili in funzione di elementi appartenenti a uno spazio dotato di prodotto interno, così da poter ragionare in termini della massimizzazione della distanza minima tra gli esempi e la superficie che separa le due classi (il cosiddetto *margine*). Più formalmente, i dati a disposizione vengono modellati utilizzando un campione etichettato avente forma $\{(x_i,\ y_i),\ i = 1,\ \ldots,\ m\} \subseteq X \times Y$, dove X indica il sopra menzionato spazio degli oggetti, $Y = \{-1,\ +1\}$ viene utilizzato per associare a ogni oggetto un'etichetta che descrive una tra due classi e m indica il numero totale di esempi, intesi come coppie (oggetto, etichetta). Soffermandosi sulla ricerca di classificatori lineari, basati cioè sulla posizione relativa degli oggetti rispetto a un iperpiano contenuto all'interno di X e avente equazione $w \cdot x + b = 0$ (per $w \in X$ e $b \in \mathbb{R}$), la sopra menzionata massimizzazione del margine si traduce nella soluzione del seguente problema di ottimizzazione vincolata:

$$\begin{aligned} \min\ & \tfrac{1}{2}\, w \cdot w + C \textstyle\sum_{i=1}^{m} \xi_i \\ & w \cdot x_i + b \geq +1 - \xi_i \text{ per ogni } i \text{ tale che } y_i = +1, \\ & w \cdot x_i + b \leq -1 + \xi_i \text{ per ogni } i \text{ tale che } y_i = -1, \\ & \xi_i \geq 0 \text{ per ogni } i = 1,\ \ldots,\ m. \end{aligned} \tag{1}$$

L'idea è quella di determinare una coppia $(w^*,\ b^*)$ tale che $w^* \cdot x + b^*$ abbia segno positivo ogni volta che x è sostituito con un oggetto avente +1 come etichetta, con un'analoga proprietà per i rimanenti oggetti. I vincoli in (1) rilassano in realtà questa richiesta, ammettendo degli errori di classificazione tramite l'introduzione di variabili *slack*. La funzione obiettivo combina due scopi contrastanti: i) massimizzare il margine della soluzione e ii) minimizzare il numero di errori; la costante $C > 0$ permette di dare un peso separato a queste due componenti durante il processo di ottimizzazione.

La soluzione di (1) viene determinata passando al suo duale, che assume la forma

$$\begin{aligned} \max\ & \textstyle\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i\, \alpha_j\, y_i\, y_j\, x_i \cdot x_j \\ & \textstyle\sum_{i=1}^{m} \alpha_i\, y_i = 0 \\ & \alpha_i \geq 0 \text{ per ogni } i = 1,\ \ldots,\ m, \\ & \alpha_i \leq C \text{ per ogni } i = 1,\ \ldots,\ m, \end{aligned} \tag{2}$$

il cui valore ottimale $\alpha_1^*, \ldots, \alpha_m^*$ per le variabili lagrangiane permette di determinare la soluzione (w^*, b^*) del problema (1):

$$\begin{aligned} w^* &= \sum_{i=1}^{m} \alpha_i^* \, y_i \, x_i, \\ b^* &= y_i - w^* \cdot x_i \text{ per qualsiasi } i \text{ tale che } 0 < \alpha_i^* < C, \end{aligned} \tag{3}$$

dove il valore per b^* si ottiene considerando alcune note relazioni di dualità note come condizioni di Karush-Kuhn-Tucker, che nella fattispecie richiedono che risulti $\alpha_i^*(y_i(w \cdot x_i + b) - 1 + \xi_i) = 0$ e $(C - \alpha_i)\,\xi_i = 0$. La classificazione di un generico oggetto x^g (che sia uno degli oggetti visti durante la fase di addestramento o che si tratti di un oggetto per il quale non è disponibile un'etichetta) avviene calcolando il segno di $w^* \cdot x^g + b^*$, o alternativamente di $\sum_{i=1}^{m} \alpha_i^* \, y_i \, x_i \cdot x^g + b^*$.

Nonostante questo modello sia essenzialmente legato all'uso di classificatori lineari, è possibile mantenerne l'impalcatura anche in presenza di dati che impongano al contrario di considerare superfici di separazione non lineari: la soluzione consiste nell'applicare a ogni oggetto $x_i \in X$ una trasformazione non lineare Φ per poi considerare il problema di trovare un classificatore lineare per gli esempi ottenuti associando l'etichetta y_i all'immagine $\Phi(x_i)$. In questo modo viene indotta su X una superficie di separazione non lineare, in funzione di come è stata scelta Φ. Si dimostra facilmente che ciò equivale a considerare al posto di (2) il problema

$$\begin{aligned} \max \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \, \alpha_j \, y_i \, y_j \, k\left(x_i, x_j\right) \\ & \sum_{i=1}^{m} \alpha_i \, y_i = 0 \\ & 0 \le \alpha_i \le C \text{ per ogni } i = 1, \ldots, m, \end{aligned} \tag{4}$$

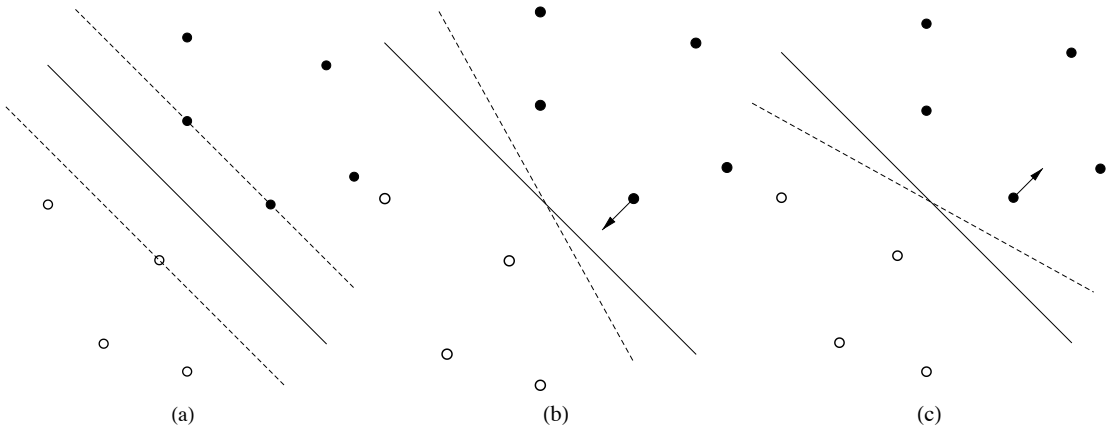
dove $k\left(x_i, x_j\right)$ indica il prodotto interno nel codominio di Φ delle immagini di x_i e x_j (più in generale k va a denotare la funzione *kernel* associata alla trasformazione Φ). Il calcolo dei parametri del classificatore corrispondente alla soluzione di questo problema è simile a quello descritto in (3), con l'importante differenza che non risulta più possibile ottenere direttamente il valore di w^*, bensì è necessario calcolare ogni prodotto $w^* \cdot x$ tramite l'espansione $\sum_{i=1}^{m} \alpha_i^* \, y_i \, k(x_i, x)$. Pertanto la classificazione di un generico oggetto x^g avviene calcolando il segno di $\sum_{i=1}^{m} \alpha_i^* \, y_i \, k(x_i, x^g) + b^*$. Questo approccio alla classificazione ha riscosso un notevole successo negli ultimi quindici anni, sia per il notevole numero di applicazioni che a causa di proprietà notevoli della soluzione proposta, tra cui si annoverano la sparsità (di norma gran parte dei valori ottimali $\alpha_1^*, \ldots, \alpha_m^*$ risultano nulli), la buona capacità di generalizzazione e la relativa immunità dal degrado delle prestazioni al crescere della dimensione di X.

Classificazione di dati di qualità variabile

Assumiamo che ogni esempio a disposizione sia descritto, oltre che da un oggetto $x_i \in X$ e da un'etichetta $y_i \in Y$, anche da un ulteriore valore q_i che descriva quantitativamente la qualità dell'esempio stesso, con la convenzione che valori positivi indicano confidenza nell'associazione dell'oggetto descritto da x_i alla classe indicata da y_i e valori negativi siano da collegare all'assenza di tale confidenza. Assumiamo anche che, in entrambi i casi, questa confidenza risulti tanto più forte quanto più grande è il valore assoluto di q_i. L'idea alla base di questo lavoro è quella di estendere l'algoritmo di classificazione descritto nel paragrafo precedente in modo da poter sfruttare la qualità dei dati disponibili durante il processo di apprendimento, così che per esempio la ricerca della soluzione nei problemi di ottimizzazione penalizzi l'errata classificazione di esempi di elevata qualità promuovendo nel contempo la contraddizione dei dati in caso di bassa qualità.

L'idea alla base di questa variante dell'algoritmo è quella di riposizionare gli oggetti all'interno di X in funzione della loro qualità, spostandoli più precisamente nella direzione indicata dalla normale all'iperpiano che separerebbe le due classi (rispettivamente in X nella versione originale dell'algoritmo e nel codominio di Φ nella variante non lineare) se si utilizzasse l'algoritmo originale. Il verso dello spostamento viene invece determinato dal segno di q_i, in modo che:

- quando un esempio è associato a un valore $q_i > 0$, e quindi quando viene considerato importante, il corrispondente oggetto x_i viene avvicinato all'iperpiano separatore del problema originale, con il risultato di allontanare la superficie di separazione nel problema modificato dall'oggetto di partenza;
- se invece $q_i < 0$, lo spostamento allontana l'oggetto dal separatore del problema originale, così che nella nuova soluzione la superficie di separazione si avvicina all'oggetto di partenza ed eventualmente lo sorpassa, rendendolo così erroneamente classificato.



(a) (b) (c)

Figura 1. La variante proposta per il problema di classificazione: ogni cerchio indica un esempio, dove la posizione individua un oggetto in $X = \mathbb{R}^2$ e il colore bianco o nero determina una delle due classi. (a) Nella formulazione originale del problema, i punti hanno una posizione prefissata rispetto all'iperpiano separatore (linea grigia), il quale viene determinato in modo da massimizzare il margine (inteso come la distanza tra le linee tratteggiate). (b) Nell'algoritmo studiato, un punto con valore positivo per q_i viene spostato vicino alla superficie di separazione del problema originale (linea grigia), come indicato dalla freccia; in questo modo la superficie di separazione effettiva (linea tratteggiata) va ad allontanarsi dalla posizione originale del punto. (c) In modo analogo, punti con valori negativi di q_i vengono allontanati dalla superficie di separazione del problema originale.

La Figura 1 visualizza un esempio di questi due tipi di spostamenti, confrontandoli con la situazione corrispondente nel problema originale.

Se lo spostamento dell'i-esimo punto viene fatto proporzionalmente al valore di q_i il problema (1) si trasforma in

$$\begin{aligned} \min \quad & \tfrac{1}{2}\, w \cdot w + C \textstyle\sum_{i=1}^{m} \xi_i \\ & w \cdot \left(x_i - \tfrac{q_i}{2}\, w \right) + b \geq +1 - \xi_i \text{ per ogni } i \text{ tale che } y_i = +1, \\ & w \cdot \left(x_i + \tfrac{q_i}{2}\, w \right) + b \leq -1 + \xi_i \text{ per ogni } i \text{ tale che } y_i = -1, \\ & \xi_i \geq 0 \text{ per ogni } i = 1, \ldots, m, \end{aligned} \tag{5}$$

e la sua formulazione duale diventa

$$\begin{aligned} \max \quad & \textstyle\sum_{i=1}^{m} \alpha_i - \frac{1}{2\left(1 - \sum_{i=1}^{m} \alpha_i q_i\right)} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ & \textstyle\sum_{i=1}^{m} \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \text{ per ogni } i = 1, \ldots, m, \\ & \alpha_i \leq C \text{ per ogni } i = 1, \ldots, m, \end{aligned} \tag{6}$$

la cui soluzione ottimale $\alpha_1^*, \ldots, \alpha_m^*$ permette di ottenere i valori che risolvono il problema originale come

$$\begin{aligned} & w^* = \frac{1}{1 - \sum_{i=1}^{m} \alpha_i q_i} \sum_{i=1}^{m} \alpha_i^* y_i x_i, \\ & b^* = y_i - w^* \cdot x_i + y_i \frac{q_i}{2}\, w^* \cdot w^* \text{ per qualsiasi } i \text{ tale che } 0 < \alpha_i^* < C. \end{aligned} \tag{7}$$

Va notato come questa formulazione perda uno dei principali vantaggi del problema originale, che consiste nel richiedere di ottimizzare una funzione obiettivo quadratica. Sperimentalmente (cfr. il paragrafo successivo) si osserva come anche questa nuova funzione obiettivo risulti ottimizzabile tramite strumenti computazionali standard quando si lavora con classificatori lineari.

Volendo invece utilizzare dei kernel per indurre delle superfici di separazione non lineari, si sperimenterebbe spesso che il problema (6), una volta modificato a seguito dell'introduzione dei kernel, non è risolubile in modo efficiente. A questo scopo è possibile considerare soluzioni subottimali, moltiplicando la funzione obiettivo per $1 - \sum_{i=1}^{m} \alpha_i q_i$ e aggiungendo nel contempo un vincolo che richieda che questa quantità risulti positiva.

Simmetrizzando il valore ottenuto per la funzione obiettivo, cioè sostituendo ogni elemento $s_{i,j}$ con $\frac{1}{2}\left(s_{i,j} + s_{j,i}\right)$ si ottiene la seguente versione del problema:

$$\begin{aligned} \max \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \left(k\left(x_i, x_j\right) - y_i y_j \left(q_i + q_j\right)\right) \\ & \sum_{i=1}^{m} \alpha_i y_i = 0 \\ & \sum_{i=1}^{m} \alpha_i q_i < 1 \\ & 0 \le \alpha_i \le C \text{ per ogni } i = 1, \ldots, m, \end{aligned} \tag{8}$$

che ha il vantaggio di avere una funzione obiettivo quadratica.

Esperimenti

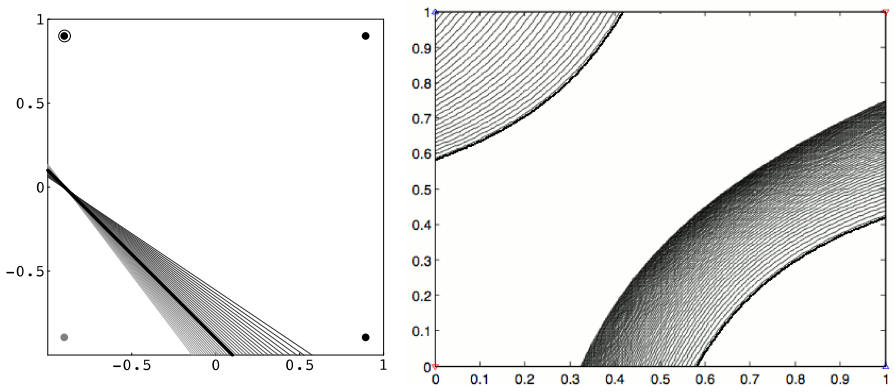
I paragrafi seguenti descrivono degli esperimenti volti all'analisi dell'algoritmo proposto tramite l'elaborazione di dati sintetici e la sua applicazione a problemi reali. Le elaborazioni si basano sul package *svMathematica* [11], costruito al fine di accoppiare Wolfram *Mathematica* con il software di modellazione ampl [6] e l'ottimizzatore numerico snopt [19].

■ Analisi di sensitività su dati sintetici

Una prima applicazione dell'algoritmo proposto è legata all'uso di dati sintetici, al fine di valutare l'impatto della scelta di valori numerici specifici per la qualità degli esempi sui risultati ottenuti. L'esperimento e i suoi risultati sono descritti in Figura 2. Si considerano quattro oggetti posizionati ai vertici di un quadrato: quello in basso a sinistra viene associato alla classe la cui etichetta è pari a -1 mentre quelli rimanenti sono associati alla classe con etichetta positiva. In questo modo i dati sono separabili da una retta e quindi è giustificato l'utilizzo della versione lineare dell'algoritmo. Il valore di qualità per tutti i punti è impostato a 0, tranne il punto in alto a sinistra che viene associato a venti diversi valori di qualità, equispaziati tra -0.5 e 0.5. Per ognuno di questi valori viene risolto il problema di classificazione, sovrapponendo in Figura 2(a) il fascio di rette risultanti al campione di partenza.

Si vede facilmente come le rette in output dall'algoritmo, che rappresentano i classificatori ottenuti, tendono a ruotare così da allontanarsi dal punto in alto a sinistra man mano che il valore di qualità aumenta. Viceversa, la rotazione avvicina rette e punto al diminuire del valore di qualità. Si noti anche come l'output dell'algoritmo proposto coincida con quello della versione originale (la linea con spessore maggiore) quando a tutti i punti viene associato un valore nullo per descrivere la qualità. La Figura 2(b) riassume i risultati di un esperimento analogo basato su superfici di separazione polinomiali.

La Figura 3 visualizza i risultati di un secondo esperimento basato su dati sintetici che coinvolge un numero non banale di esempi. Più precisamente, sono stati considerati 150 punti estratti uniformemente a caso in $[0, 1]^2$ ed etichettati rispetto a una curva prefissata ℓ.



(a) (b)

Figura 2. (a) Analisi di sensitività per l'algoritmo proposto in forma lineare, al variare della qualità del punto in alto a sinistra (evidenziato con un cerchio esterno, mentre grigio e nero indicano le due diverse classi) in un insieme di venti valori equispaziati tra -0.5 e 0.5. Per ogni valore si ottiene una retta che separa le classi, utilizzata per costruire il fascio disegnato in figura. Il fascio contiene anche la retta corrispondente alla soluzione della versione originale dell'algoritmo, che non sfrutta i valori di qualità, e che è stata disegnata utilizzando uno spessore maggiore. (b) Un esperimento analogo basato su superfici di separazione polinomiali.

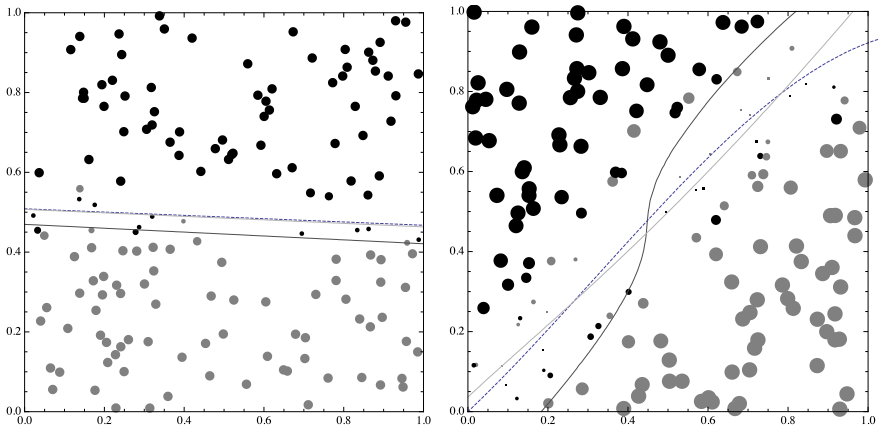


Figura 3. Esperimenti di ricostruzione a partire da campioni etichettati da (a) linee e (b) polinomi di terzo grado. I cerchi rappresentano punti da dividere in due classi, caratterizzate dai colori nero e grigio. Il raggio di ogni cerchio è proporzionale al suo valore di qualità. Le rette e le curve tratteggiate, grigie e grigio chiare rappresentano le superfici rispettivamente utilizzate per effettuare l'etichettatura iniziale dei punti e quelle ottenute in output dall'algoritmo proposto e da quello originale.

Per ogni punto x_i è poi stata calcolata la distanza $d(x_i)$ rispetto a ℓ e:

- l'etichetta del punto è stata scambiata con quella della rimanente classe con probabilità pari a $e^{-\alpha\, d(x_i)^2}$, al fine di simulare degli errori di misurazione;
- al punto è stato associato il valore di qualità $q_i = 1 - e^{-\alpha\, d(x_i)^2} - \delta$, così da legare la qualità dei dati alla probabilità che questi siano affetti da errori di misurazione.

Nei due punti precedenti, α e δ indicano valori scelti opportunamente sulla base di tentativi. In questo modo, i punti vicini a ℓ hanno maggior probabilità di vedere modificata la loro etichetta, e contemporaneamente tali punti si vedono assegnato un valore basso di qualità, così che il processo di apprendimento sia portato a selezionare non la classe (sbagliata) indicata dalla loro etichetta, bensì l'altra, che incidentalmente è quella corretta. La Figura 3 visualizza due diversi campioni dopo la procedura di ri-etichettatura (in cui ℓ risulta essere rispettivamente una retta e un polinomio di terzo grado) sopra descritta: ogni cerchio rappresenta un punto, il cui colore identifica la classe a cui è stato assegnato e il cui raggio è scelto in funzione del suo valore di qualità. La figura indica anche la curva utilizzata per etichettare inizialmente i punti (tracciata con stile tratteggiato) e le superfici di separazione ottenute applicando l'algoritmo proposto e la versione originale (tracciate rispettivamente in grigio e in grigio chiaro). In entrambi i casi si vede come il risultato dell'algoritmo proposto rappresenti un'approssimazione decisamente migliore della curva di partenza rispetto a quello dell'algoritmo originale. In particolare, le due superfici si sovrappongono nel caso lineare.

■ Riconoscimento di cifre manoscritte

Un ambito in cui è possibile ottenere facilmente dei dati di qualità sensibilmente differente è quello del riconoscimento delle cifre manoscritte. Partendo da un sottoinsieme di duemila immagini che descrivono siffatte cifre all'interno del database MNIST [23] già diviso in training e test set e selezionando quelle associate alle cifre 0 e 1, si sono ottenuti un training e un test set rispettivamente di 213 e 211 elementi. Per poter associare dei valori di qualità si è seguita la procedura descritta in [12], essenzialmente basata sulla divisione in un numero fisso di *cluster*, ottenuti con l'algoritmo di *fuzzy c-means* [1], calcolando il valore di qualità di un esempio come massimo grado di appartenenza del punto corrispondente ai vari *cluster*, così che

$$q_i = \max_{1 \le k \le c} \left\{ \left(\sum_{j=1}^{c} \left(\frac{\| x_i - v_k \|}{\| x_i - v_j \|} \right)^{\frac{2}{\alpha - 1}} \right)^{-1} \right\}, \tag{9}$$

dove c e v_i indicano rispettivamente il numero di *cluster* e il centro dell'i-esimo *cluster*. A titolo di esempio, la Figura 4 visualizza i 10 migliori e i 10 peggiori rappresentanti della cifra 1.

(a) (b)

Figura 4. I dieci migliori (a) e i dieci peggiori (b) rappresentanti della cifra 1 identificati dalla procedura di clustering.

I risultati ottenuti sono riepilogati in Tabella 1: essenzialmente si ottiene un'accuratezza completa sia in fase di addestramento, sia in fase di test. I risultati sono meno chiari quando si cerca di discriminare tra i 3 e gli 8, perché in questo caso la procedura di divisione tende a sovrapporre i vari *cluster*. Per ovviare a questa limitazione è stato selezionato un sottoinsieme dei dati a disposizione i cui valori di qualità sono stati impostati manualmente a 1 e a -1 sulla base di un'analisi visuale, mentre gli esempi non esaminati sono stati associati a un valore nullo di qualità. Il risultato ottenuto, sempre riepilogato in Tabella 1, è meno interessante dell'altra parte dell'esperimento ma comunque confrontabile con gli altri approcci descritti in letteratura [9], basati però su superfici di separazione non lineari accoppiate a procedure di *pre-processing* più articolate.

Data set	Train	Test
0/1	100 %	100 %
3/8	100 %	90 %

Tabella 1. Risultati della procedura di riconoscimento di cifre manoscritte, espressi in termini di accuratezza della classificazione in fase di addestramento (Train) e di validazione (Test) dell'esperimento. Ogni riga fa riferimento a un esperimento volto a discriminare tra due particolari cifre, indicate in corrispondenza della colonna "Data set".

■ Classificazione di proteine sulla base di annotazioni

Un ambito caratterizzato dall'esplicita presenza di esempi accoppiati a un'esplicita valutazione della loro importanza è quello della classificazione dei geni e della predizione automatica della loro classe funzionale. A causa dell'estrema complessità dei fenomeni chimici e fisici alla base dei processi biologici, l'annotazione funzionale dei geni (depositari dell'informazione necessaria alla sintesi dei costituenti di base dei viventi, le proteine) rappresenta una sfida fondamentale in biologia computazionale.

Le conoscenze attualmente disponibili rispetto ai processi biologici sono state recentemente strutturate nella forma di complesse ontologie gerarchiche composte da centinaia o migliaia di termini e costituite da una foresta di alberi (MIPS, Functional Catalogue FunCAT [17]) o da grafi diretti aciclici (Gene Ontology, GO [8]). Tale rappresentazione è necessaria in quanto i processi biologici sono estremamente difficili da definire come entità a sé stanti (un caso tipico è costituito dal metabolismo dei carboidrati, realizzato mediante diverse vie metaboliche interconnesse). Una volta definita l'ontologia funzionale, degli esperti in un team associano i geni presenti nel DNA di un organismo di interesse con i termini dell'ontologia funzionale. A questo riguardo è importante notare che il processo di annotazione è realizzato dagli esperti in modo indipendente e, quindi, non è infrequente che il medesimo gene sia associato ai termini dell'ontologia mediante "evidenze" di qualità differenti.

Un ipotetico gene g_1 potrebbe essere associato al termine (classe funzionale) T_1 a causa dell'esito di un esperimento di laboratorio (evidenza forte), mentre un secondo gene g_2 potrebbe essere associato (annotato) al termine T_1 a causa dell'ipotesi formulata da un autore nelle conclusioni di un lavoro descritto in letteratura, pur senza avere un supporto sperimentale (evidenza debole). È facile intuire come, nel caso appena descritto, in un ipotetico task di predizione della funzione dei geni l'etichetta T_1 associata al gene g_1 debba essere considerata più "sicura" (o di qualità maggiore) rispetto alla medesima etichetta associata al gene g_2. Il problema è complicato dal fatto che il processo di annotazione è dinamico: le annotazioni funzionali dei geni vengono costantemente aggiornate sulla base dei dati pubblicati in letteratura. Questo determina l'esistenza, per ogni coppia gene-termine funzionale, di numerose annotazioni, ognuna avente un grado di qualità differente.

Nell'esperimento presentato il problema è stato risolto esaminando, per ogni gene e ogni termine funzionale, la collezione di evidenze presente nel database GO, e scegliendo come evidenza di associazione tra gene e termine quella più forte tra quelle disponibili. Il ranking di qualità delle annotazioni è stato effettuato utilizzando le regole di conversione descritte in Tabella 2, pubblicate in letteratura [2].

Codice	Definizione	Ranking
IDA	Inferred from direct assay	5
IGI	Inferred from genetic interaction	5
IMP	Inferred from mutant phenotype	5
IPI	Inferred from physical interaction	5
IC	Inferred from curator	4
TAS	Traceable author statement	4
IEP	Inferred from expression pattern	3
RCA	Inferred from reviewed computational analysis	3
IGC	Inferred from genomic context	3
ISS	Inferred from sequence or structural similarity	2
IEA	Inferred from electronic annotation	2
NAS	Non − traceable author statement	2
NR	Not recorded	1
ND	No biological data available	0

Tabella 2. Corrispondenza tra etichette categoriche che descrivono le associazioni e valori numerici di qualità.

Pertanto, se nell'esperimento descritto il gene g_1 è annotato con il termine funzionale T_1 mediante il set di evidenze TAS (4), IEP (3), NAS (2) e IDA (5), verrà considerato associato al termine T_1 con il peso assoluto dell'evidenza più forte, cioè 5 a causa della presenza di un'annotazione sperimentale di tipo IDA (inferred from direct assay). I due task di classificazione proposti sono stati progettati in modo da coinvolgere un numero confrontabile di geni. Essi si basano su un set di esempi (geni) positivi associati al termine GO:0006457 (ontology_type: biological process, description: protein folding), mentre il set dei geni negativi varia nei due task. In particolare nel task 1, tale insieme è costituito dai geni annotati con il termine GO:0051704 (ontology_type: biological process, description: multi-organism process), mentre nel task 2 gli esempi negativi sono geni annotati con il termine GO:0006468 (ontology_type: biological process, description: protein amino acid phosphorylation).

La scelta dei due termini utilizzati per definire i set degli esempi negativi è basata unicamente su due criteri: numero confrontabile di geni rispetto al termine cui sono associati i geni che costituiscono gli esempi positivi in entrambi i task di classificazione e similarità tra i termini dell'ontologia funzionale considerata. Per il calcolo delle similarità tra i termini dell'ontologia GO sono state utilizzate le funzioni messe a disposizioni dal package R *GOSim* [7], utilizzando la distanza di Lin [10].

Dato che la similarità tra i termini coinvolti nel task 1 è minore di quella tra i termini coinvolti nel task 2 è lecito considerare più complesso il task 2 rispetto al task di classificazione 1. L'esperimento è stato concepito per preservare il più possibile le caratteristiche dei due problemi di classificazione (frequenza degli esempi associati alle diverse classi di evidenza — cfr. Tabella 2 — e numero degli esempi coinvolti), facendo variare unicamente la similarità tra i termini funzionali dell'ontologia.

Il dataset di input utilizzato per gli esperimenti di classificazione è costituito da interazioni tra proteine ottenute dal database BioGRID [20]. In questo database sono riportate le interazioni fisiche e genetiche tra proteine riportate in letteratura. L'utilizzo di questo tipo di dati è giustificato dall'attesa che se due proteine cooperano alla realizzazione di un medesimo processo biologico hanno più probabilità di interagire rispetto a proteine coinvolte in processi diversi, in quanto le proteine che partecipano ai medesimi processi hanno una maggiore probabilità di trovarsi in compartimenti cellulari dedicati in modo specifico alla realizzazione dei processi stessi. I dati di interazione proteica hanno avuto un ruolo fondamentale nel processo di annotazione delle funzioni geniche, e il loro valore può essere apprezzato notando (cfr. Tabella 2) che gli score di evidenza ad essi associati (IPI ed IGI) sono i più alti tra quelli disponibili. Gli esperimenti sono basati sulla classificazione di geni del lievito (*Saccharomices cerevisiae*). La numerosità dei campioni di entrambi i task di classificazione è riportata in Tabella 3, specificando il numero complessivo di geni e il numero di esempi positivi e negativi.

L'esperimento è stato condotto nel modo seguente: preventivamente sono state ordinate le etichette che descrivono la qualità delle annotazioni, così da poter assegnare a ognuna di esse un valore numerico, come indicato in Tabella 2. I valori di qualità q_i dei vari esempi sono successivamente calcolati dividendo tali numeri per il valore di un ulteriore parametro θ. Il metodo proposto è quindi stato applicato considerando due insiemi di valori per i parametri C e θ (rispettivamente $\{10^i,\ i = 2,\ 3,\ \ ..., 6\}$ e $\{i\,10^2,\ i = 2,\ ..., 12\}$), estraendo tutte le possibili combinazioni e addestrando per ognuna di esse un classificatore.

	N. geni	**N. esempi positivi**	**N. esempi negativi**
Task1	235	99	136
Task2	240	101	139

Tabella 3. Numerosità dei campioni utilizzati negli esperimenti.

Più precisamente, per ogni data combinazione i dati a disposizione sono stati partizionati casualmente in due insiemi contenenti rispettivamente il 60% e il 40% degli esempi annotati, usando il primo insieme per addestrare un classificatore e il secondo per valutarne la *performance*, in termini di opportuni indici (cfr. più avanti). Questa operazione è stata ripetuta dieci volte, modificando ogni volta la ripartizione tra i due insiemi di dati, e calcolando il valore medio degli indici di *performance*. Alla fine della procedura è stato selezionato il classificatore con la migliore *performance* media.

Al fine di poter confrontare i risultati ottenuti con lo stato dell'arte, un esperimento simile è stato eseguito utilizzando però la versione originale dell'algoritmo, applicando una procedura di *model selection* per il parametro *C* considerando tutti i valori nell'insieme $\{2^i,\ i = -5,\ -3,\ ...,\ 15\}$. Una volta determinato il valore ottimale per tale parametro, in analogia con quanto visto sopra sono stati addestrati dieci classificatori, ogni volta ripartendo casualmente i dati a disposizione in due insiemi contenenti rispettivamente il 60% e il 40% degli esempi, addestrando sul primo insieme e valutando sul secondo, riportando poi la media delle *performance*.

Gli indici utilizzati per valutare la bontà dei classificatori ottenuti sono due:

- l'*errore assoluto*, inteso come frequenza di esempi erroneamente classificati;
- l'*errore pesato*, in cui la frequenza descritta al punto precedente è pesata rispetto ai valori numerici associati alle annotazioni in Tabella 2.

In sintesi, l'errore assoluto conta semplicemente il numero di esempi che non vengono correttamente descritti dal classificatore, indipendentemente dalla loro importanza, mentre l'errore pesato tenderà a essere più alto nel caso in cui il classificatore selezionato non descriva bene degli esempi importanti e viceversa. La Tabella 4 riporta i valori di tali indici per gli esperimenti basati sia sul metodo proposto che per quello classico, indicati rispettivamente con qSVC e SVC. Si nota come il metodo proposto consegua dei risultati migliori in entrambi i task relativamente all'errore pesato.

	Task1		**Task2**	
	SVC	**qSVC**	**SVC**	**qSVC**
Errore assoluto	22.24 %	19.36 %	22.13 %	22.23 %
Errore pesato	20.36 %	17.35 %	19.20 %	16.95 %

Tabella 4. Risultati della classificazione di proteine.

Conclusioni

È stata presentata una famiglia di algoritmi di classificazione basati su vettori di supporto caratterizzati dalla capacità di considerare ogni esempio utilizzato in fase di addestramento sulla base della sua importanza relativa. L'approccio descritto, motivato descrivendo alcuni campi applicativi caratterizzati da informazione di qualità eterogenea, è stato applicato sia a dati prodotti sinteticamente al fine di valutare la sensitività dell'algoritmo rispetto ai parametri che lo caratterizzano sia a problemi reali tratti dai campi applicativi sopra menzionati.

Bibliografia

[1] Bezedek, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981

[2] Buza, T. J., McCarthy, F. M., Wang, N., Bridges, S. M. and Burgess, S. C., Gene ontology annotation quality in model eukaryotes, *Nucleic Acids Research* **36** - 2 (2008), e12

[3] Cawley, G. C., Talbot, N. L. C., Foxall, R. J., Dorling, S. R. and Mandic, D. P., Heteroscedastic kernel ridge regression, *Neurocomputing* **57** (2004), 105-124

[4] Cortes, C. and Vapnik, V., Support-vector networks, *Machine Learning* **20** (1995), 121-167

[5] Feltham, G., The value of information, *Accounting Review* **43** - 4 (1968), 684-696

[6] Fourer, R., Gay, D. M. and Kernighan, B. W., *AMPL: A Modeling Language for Mathematical Programming*, Duxbury Press/Brooks/Cole Publishing Company, 2002

[7] Fröhlich, H., Speer, N., Poustka, A. and Beißbarth, T., GOSim — an R package for computation of information theoretic GO similarities between terms and gene products, *BMC Bioinformatics* **8** (2007), 166

[8] The Gene Ontology Consortium, The Gene Ontology (GO) database and informatics resource, *Nucleic Acids Research* **32** (2004), D258-D261

[9] LeCun, Y., Bottou, L. Bengio, Y. and Haffner, P., Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86** (November 1998), 2278-2324

[10] Lin, D., An information-theoretic definition of similarity, In Kaufmann, M. (Ed.), *Proceedings of the 15th International Conference on Machine Learning*, volume 1, San Francisco, CA, 2007, 296-304

[11] Malchiodi, D. and Valerio, L., svMathematica: implementazione in *Mathematica* di algoritmi di machine learning basati su vettori di supporto, in *Atti del Mathematica Italia User Group Meeting*, Adalta, 2009

[12] Malchiodi, D., Embedding sample points uncertainty measures in learning algorithms, *Nonlinear Analysis: Hybrid Systems* **2** (2008), 635-647

[13] Marubini, E. and Valsecchi, M. G., *Analyzing Survival Data from Clinical Trials and Observational Studies*, John Wiley & Sons, Chichester, 1995

[14] Naylor, J. C. and Domine, K. D., Inferences based on uncertain data: Some experiments on the role of slope magnitude, instructions, and stimulus distribution shape on the learning of contingency relationships, *Organizational Behavior and Human Performance* **27** (1981), 1-31

[15] Patterson, K.E. and Baddeley, A. D., When face recognition fails, *Journal of experimental psychology. Human learning and memory* **3** - 4 (1977), 406 - 417

[16] Resnick, P. and Varian, H. R., Recommender systems, *Communications of the ACM* **40** - 3 (1997), 56-58

[17] Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M. and Mewes H. W., The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucleic Acids Research* **32** - 18 (2004), 5539-5545

[18] Salomon, G., The differential investment of mental effort in learning from different sources, *Educational Psychologist* **18** - 1 (1983), 42-50

[19] Stanford Business Office Inc., SNOPT 6.0, risorsa disponibile on-line, *http://www.sbsi-sol-optimize.com/asp/sol_product_snopt.htm*

[20] Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M., BioGRID: a general repository for interaction datasets, *Nucleic Acids Research* **34** (Database issue, 2006), D535-D539, *http://thebiogrid.org*

[21] Theodoridis, S. and Koutroumbas, K., *Pattern recognition*, Elsevier/Academic Press, Amsterdam, Boston, 2006

[22] Wand, Y. and Wang, R., Anchoring data quality dimensions in ontological foundations, *Communications of the ACM* **39** (1996), 86-95

[23] Wang, H. and Bengio, S., *The MNIST database of handwritten upper-case lettters*. Technical Report 04, IDIAP (2004)

[24] Wang, R. and Strong, D., Beyond accuracy: what data quality means to data consumers, *Journal of Management Information Systems* **12** (1996), 5-34