Metodi di Classificazione Basati sulla Verifica d'ipotesi: Variazioni, Estensioni e Applicazioni

Laureando: Giovanni Corradini
Relatore: Prof. Livio Finos

Dipartimento di Scienze Statistiche
Università degli Studi di Padova

26 Marzo 2021

Introduzione

- Nel 2007 Liao e Akritas hanno sviluppato un metodo di classificazione chiamato *Test Based Classification* (**TBC**), basato sulla struttura della verifica d'ipotesi e che ha molte somiglianze con l'**LDA**.
- Per estendere la TBC al contesto con $p > n$ sono stati proposti vari metodi: il più recente è l'*Instance Based Classification* (**IBC**), proposto da He et al. nel 2019.
- In questa tesi è stato proposto un metodo *test-based*, denominato *High Dimensional Test Based Classification* (**HDTBC**) che, a differenza della IBC, tiene anche in considerazione la correlazione presente tra i predittori.
- L'HDTBC, la IBC e altri modelli di classificazione binaria sono stati confrontati su **20 dataset reali**, di cui 10 con $n > p$ e 10 con $p > n$.

Classificazione Binaria

- Sia $x^{tr} = (x_0, x_1)$ l'insieme dei dati di stima, dove con $x_0 = (x_{01}, x_{02}, ..., x_{0n_0})$ e $x_1 = (x_{11}, x_{12}, ..., x_{1n_1})$ si indicano due campioni casuali generati da X_0 e da X_1 rispettivamente, e sia x^{ts} una nuova osservazione generata o da X_0 o da X_1.
- L'obiettivo della classificazione binaria è quello di assegare x^{ts} ad una delle due popolazioni (o classi) relative a X_0 e a X_1 (indicate con 0 e 1), utilizzando le informazioni contenute in x^{tr}.

Test Based Classification (TBC)

Nel 2007 Liao e Akritas hanno proposto un metodo di classificazione basato sui test d'ipotesi, chiamato **Test Based Classification**. Fissata un'ipotesi nulla (H_0), una alternativa (H_1) e un test (τ) per la verifica di H_0 contro H_1, l'algoritmo della TBC si schematizza in 3 passi:

1. Si effettua il test τ, per la verifica di H_0 contro H_1, tra (x_0, x^{ts}) e x_1 e si ottiene α_0^{oss};
2. Si effettua il test τ, per la verifica di H_0 contro H_1, tra x_0 e (x^{ts}, x_1) e si ottiene α_1^{oss};
3. Se $\alpha_0^{oss} < \alpha_1^{oss}$ allora si assegna x^{ts} alla classe 0, altrimenti si assegna x^{ts} alla classe 1.

Test Based Classification (TBC)

Il test dovrebbe fornire un livello di significatività osservato inferiore quando si alloca x^{ts} nella classe corretta. Questo, nel caso in cui si utilizzino il t-test o la T^2 di Hotelling, è motivato dal fatto che, quando si alloca x^{ts} nella classe sbagliata:

1. La differenza delle medie fra i due gruppi dovrebbe essere inferiore rispetto a quando si alloca x^{ts} nella classe corretta;
2. La varianza campionaria all'interno di questa classe dovrebbe essere superiore rispetto a quando si alloca x^{ts} nella classe corretta.

TBC vs LDA

Con classi bilanciate una versione semplificata della TBC basata sulla T^2 di Hotelling coincide con l'Analisi Discriminante Lineare.

Con $X_0 \sim N_p(\mu_0, \Sigma)$ e $X_1 \sim N_p(\mu_1, \Sigma)$, indicando con $\hat{\mu}_0$ e $\hat{\mu}_1$ i vettori delle medie campionarie di x_0 e di x_1 e con S la matrice di varianza e covarianza pooled, le T^2 di Hotelling semplificate sono pari a:

$$T_0 = \left[\frac{n\hat{\mu}_0 + x^{ts}}{n+1} - \hat{\mu}_1\right]' \left[\left(\frac{1}{n+1} + \frac{1}{n}\right) S\right]^{-1} \left[\frac{n\hat{\mu}_0 + x^{ts}}{n+1} - \hat{\mu}_1\right] \quad \text{e}$$

$$T_1 = \left[\hat{\mu}_0 - \frac{n\hat{\mu}_1 + x^{ts}}{n+1}\right]' \left[\left(\frac{1}{n} + \frac{1}{n+1}\right) S\right]^{-1} \left[\hat{\mu}_0 - \frac{n\hat{\mu}_1 + x^{ts}}{n+1}\right].$$

La **regola di classificazione della TBC semplificata** risulta essere:

$$\begin{aligned} \alpha_0^{oss} < \alpha_1^{oss} &\Leftrightarrow T_0 > T_1 \\ &\Leftrightarrow [n(\hat{\mu}_0 - \hat{\mu}_1) + (x^{ts} - \hat{\mu}_1)]' \; S^{-1} \; [n(\hat{\mu}_0 - \hat{\mu}_1) + (x^{ts} - \hat{\mu}_1)] \\ &> [n(\hat{\mu}_0 - \hat{\mu}_1) + (\hat{\mu}_0 - x^{ts})]' \; S^{-1} \; [n(\hat{\mu}_0 - \hat{\mu}_1) + (\hat{\mu}_0 - x^{ts})]. \end{aligned}$$

TBC vs LDA

L'ultima disequazione può essere riscritta come:

$$(\hat{\mu}_0 - \hat{\mu}_1)' S^{-1} x^{ts} - \frac{1}{2} (\hat{\mu}_0 - \hat{\mu}_1)' S^{-1} (\hat{\mu}_0 + \hat{\mu}_1) > 0,$$

che è la **regola di classificazione dell'LDA in caso di classi bilanciate**.

- L'equivalenza fra la TBC semplificata basata sulla T^2 di Hotelling e l'LDA non rimane valida in caso di classi non perfettamente bilanciate.
- Nella sua versione non semplificata, la **TBC** può essere vista come una **sequenza di coppie di LDA**, dove si alloca x^{ts} nella classe relativa all'α^{oss} più piccolo.

Instance Based Learning

- Nell'LDA viene stimata una sola regola di classificazione, mentre nella TBC ne viene stimata una diversa per ogni osservazione dell'insieme di verifica.
- La maggior parte dei modelli opera come l'LDA (e.g. random forest, GLM, ...): questa categoria si chiama **model based**.
- I modelli che operano come la TBC vengono invece detti **instance based**: il modello instance based più famoso è il *KNN*.
- La TBC potrebbe rivelarsi più efficace dell'LDA in termini di capacità previsiva, specialmente con poche osservazioni.
- Il tempo che occorre alla TBC per effettuare una previsione su un insieme di k osservazioni è approssimativamente $2k$ volte il tempo che impiega l'LDA.

Instance Based Classification (IBC)

- La più recente evoluzione della TBC, denominata **Instance Based Classification (IBC)** e sviluppata nel 2019 da He et al., può anche essere applicata con $p > n$; per fare ciò vengono sostituiti x_0 e x_1 con i vettori delle distanze euclidee tra x_0 e x^{ts} e tra x_1 e x^{ts}.
- Viene inoltre effettuato un passo antecedente ai tre passi dell'algoritmo della TBC che consiste nel **filtrare le osservazioni tramite KNN**, mantenendo quindi solamente i k vicini più vicini di x_0 a x^{ts} e di x_1 a x^{ts}.
- Il test utilizzato per la regola di classificazione è quello dei ranghi con segno di Wilcoxon.

High Dimensional Test Based Classification (HDTBC)

- Con lo scopo di superare il limite della IBC, che per essere utilizzata con $p > n$ utilizza distanze euclidee senza stimare (né invertire) la matrice di varianza e covarianza, viene proposto un metodo chiamato **High Dimensional Test Based Classification (HDTBC)**.
- Il metodo consiste nell'utilizzare uno stimatore per la matrice di varianza e covarianza basato su uno **shrinkage**, proposto nel 2005 da Schäfer e Strimmer, che ne garantisce l'invertibilità anche con $p > n$.
- Vengono utilizzati due stimatori diversi: uno per le varianze e uno per le covarianze, entrambi ottenuti **minimizzando analiticamente** il rispettivo **errore quadratico medio** e rendendo quindi il metodo molto efficiente a livello computazionale.

High Dimensional Test Based Classification (HDTBC)

Indicando con $\sigma^2 = (\sigma_1^2, ..., \sigma_p^2)$ le varianze dei singoli predittori, con σ_m^2 la mediana di σ^2, con s uno stimatore di σ^2 e con s_m la mediana di s, lo stimatore della j-esima componente di σ^2 dato dallo shrinking tra s_j e s_m è:

$$s_j^\star = \lambda_v s_m + (1 - \lambda_v) s_j, \ \ con \ \lambda_v \in [0, 1].$$

Indicando con $\bar{x_j} = n^{-1} \sum_{i=1}^n x_{ij}$, con $w_{ij} = (x_{ij} - \bar{x_j})^2$, con $\bar{w_j} = n^{-1} \sum_{i=1}^n w_{ij}$, con $s_j = \frac{n}{n-1} \bar{w_j}$ (l'usuale stimatore non distorto per la varianza di x_j) e con $Var(s_j) = \frac{n}{(n-1)^3} \sum_{i=1}^n (w_{ij} - \bar{w_j})^2$ e assumendo l'esistenza dei primi due momenti di s e s_m, il λ_v che minimizza

$$EQM(s^\star) = E[\sum_{j=1}^p (s_j^\star - \sigma_j^2)^2] \text{ è : } \ \lambda_v^\star = \sum_{j=1}^p Var(s_j) / \sum_{j=1}^p (s_m - s_j)^2.$$

High Dimensional Test Based Classification (HDTBC)

Indicando con Σ la matrice di varianza e covarianza di x^{tr} (la cui diagonale è composta dagli elementi di σ^2), con S uno stimatore di Σ e con T una matrice diagonale tale che $diag(T) = diag(S)$, lo stimatore delle componenti fuori dalla diagonale di Σ dato dallo shrinking fra S e T è:

$$S^\star = \lambda_c T + (1 - \lambda_c)S, \ con \ \lambda_c \in [0, 1].$$

il λ_c che minimizza $EQM(S^\star) = E(||S^\star - \Sigma||_F^2)$ è:

$$\lambda_c^\star = \sum_{j \neq i} Var(s_{ij}) / \sum_{j \neq i} s_{ij}^2.$$

La relazione presente tra la TBC e l'LDA è la stessa che sussiste tra la HDTBC e la *Shrinkage Discriminant Analysis* (*SDA*).

Dataset Utilizzati

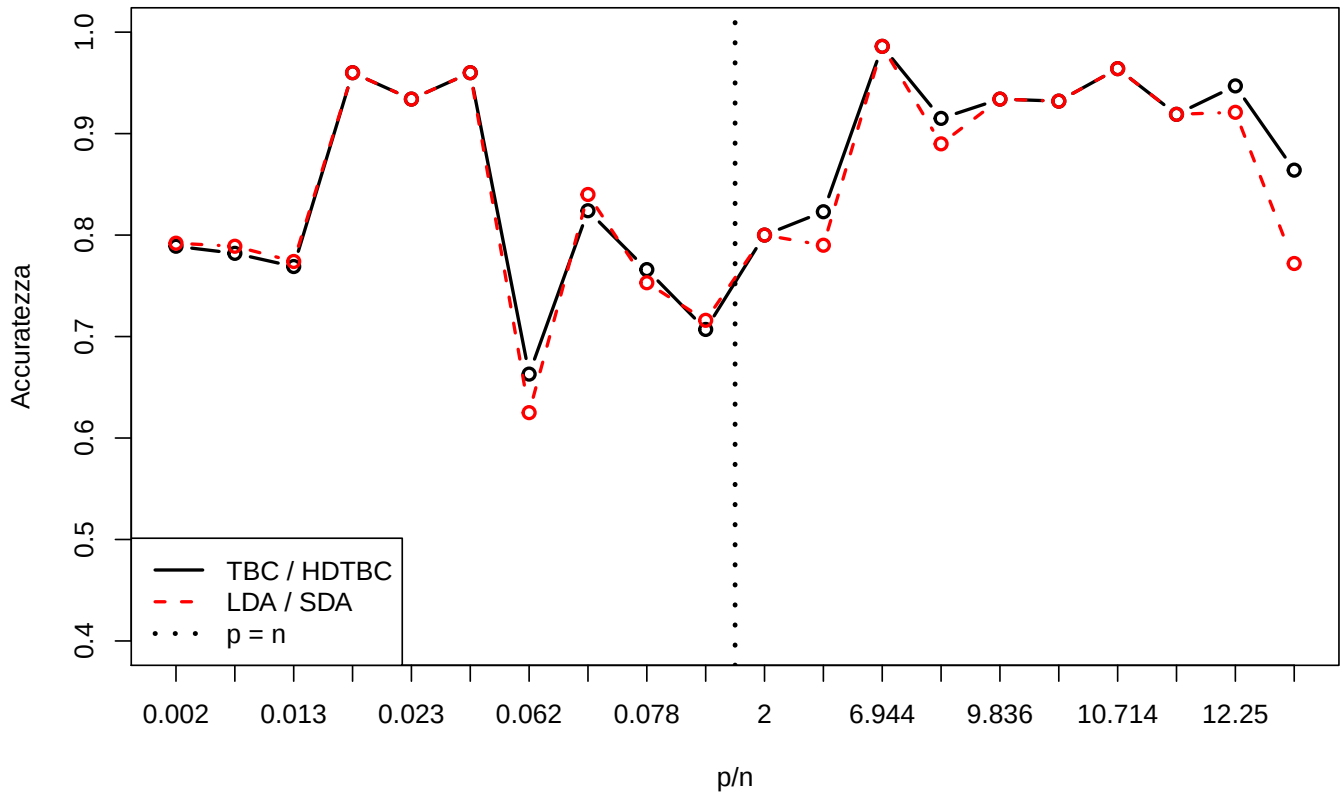
- Sono state misurate le performance predittive di vari modelli, in termini di accuratezza calcolata tramite convalida incrociata, su **20 dataset reali**, di cui 10 con $n > p$ e 10 con $p > n$.
- Nella fase di **preprocessing** sono state eliminate le osservazioni con valori mancanti, i predittori sono stati normalizzati tra $[0, 1]$ quando necessario e, per i dataset derivanti dal contesto del sequenziamento dei geni, è stato effettuato un filtraggio dei geni per mantenere solamente quelli maggiormente espressi.

Modelli Confrontati

- Oltre alla **TBC**, all'**LDA**, all'**IBC**, all'**HDTBC** e all'**SDA** sono stati confrontati anche la **Random Forest**, il **Support Vector Machine** con kernel lineare e il **GLM binomiale** con legame logistico e penalizzazione *Elastic Net*.
- Gli ultimi tre modelli sono stati adattati tramite il pacchetto *caret* che ha permesso un'**ottimizzazione degli iperparametri** tramite covalida incrociata molto efficiente.
- Il numero di vicini più vicini della IBC è stato selezionato tramite convalida incrociata.

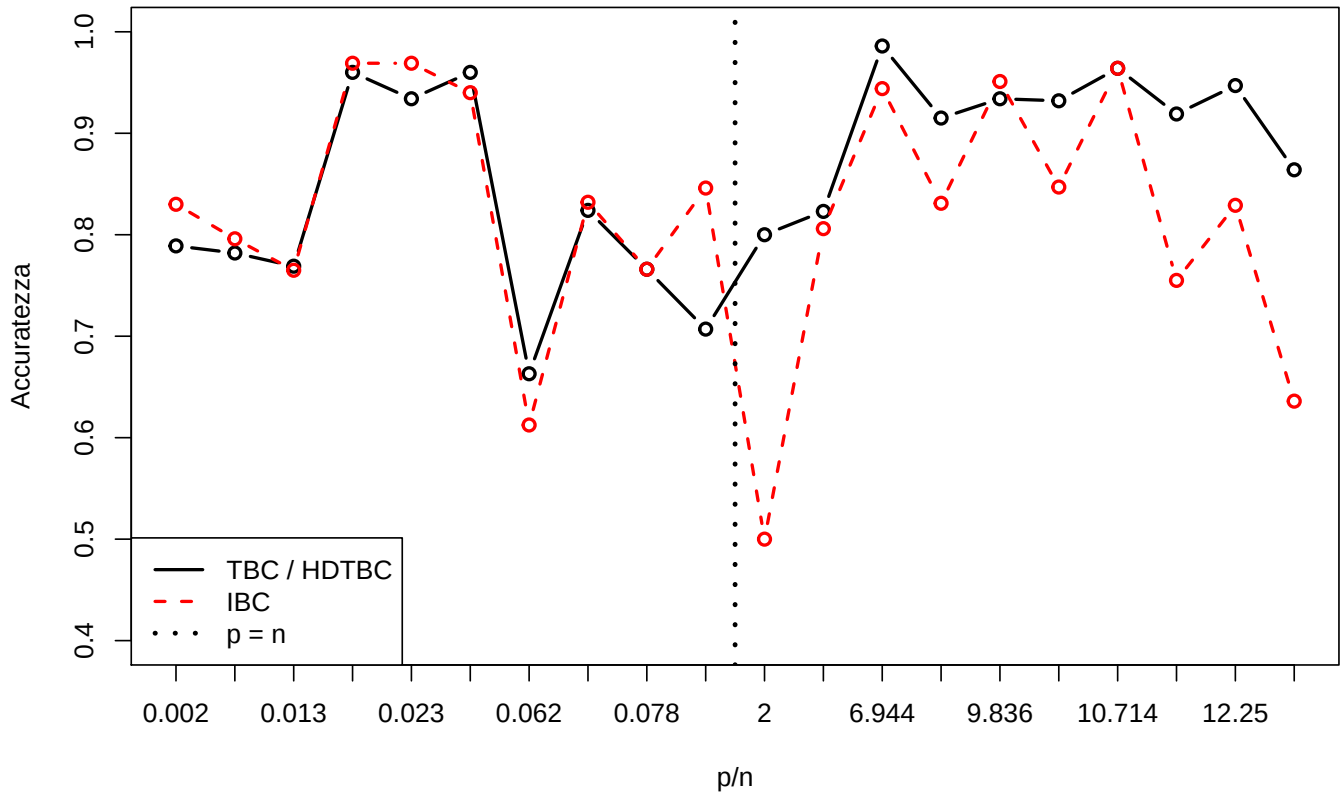
Risultati Ottenuti

A sinistra della linea verticale punteggiata viene indicata l'accuratezza di TBC e LDA (per $n > p$) sui vari dataset analizzati; a destra della linea viene invece indicata l'accuratezza di HDTBC e SDA (per $p > n$).



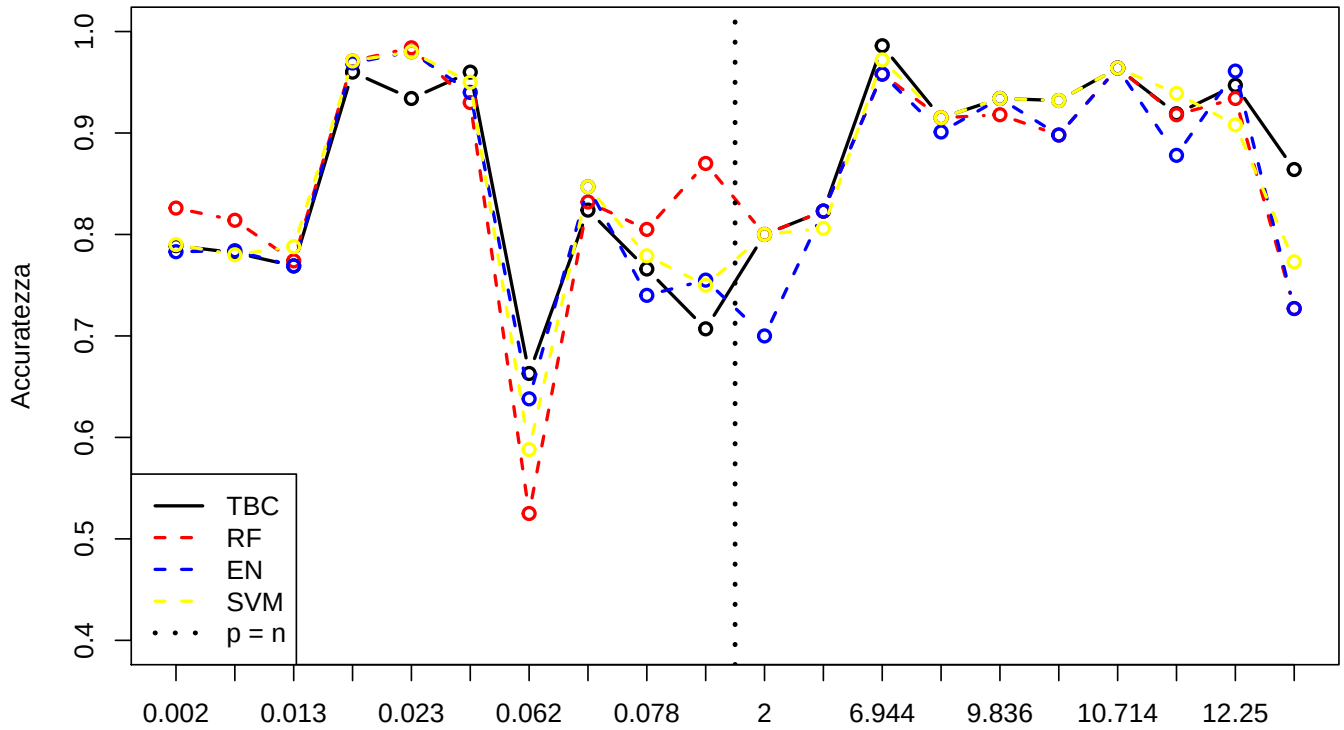
Risultati Ottenuti

Accuratezza di TBC (HDTBC con $p > n$) e di Instance Based Classification (IBC) sui vari dataset analizzati, ordinati in base al rapporto tra il numero di predittori e il numero di osservazioni.



Risultati Ottenuti

Accuratezza di TBC (HDTBC con $p > n$), Random Forest (RF), GLM binomiale con penalizzazione Elastic Net (EN) e Support Vector Machine (SVM) sui vari dataset analizzati, ordinati in base al rapporto tra il numero di predittori e il numero di osservazioni.



Conclusione

- Sui 10 dataset con $p > n$ analizzati, le **capacità previsive** della HDTBC si sono dimostrate **superiori** rispetto a quelle della IBC.
- Sono emerse forti **affinità** tra la TBC e l'LDA nel caso con $n > p$, e tra la HDTBC e l'SDA nel caso con $p > n$.
- Nel caso con $p > n$, la HDTBC sembra essere un metodo di classificazione **competitivo** con alcuni tra i metodi più famosi, sia in termini di efficacia che in termini di efficienza.

Spunti Futuri

- Modificare la TBC per tenere in considerazione contesti con **classi fortemente sbilanciate**: una possibile soluzione per gestire questo problema potrebbe consistere nell'assegnare x^{ts} alla classe 0 se $(1-\ell)\ \alpha_0^{oss} < \ell\ \alpha_1^{oss}$, con $\ell \in [0,1]$.
- Estendere la TBC alla **multiclassificazione**: un possibile metodo per operare con più di due classi consiste in una TBC *one vs all* iterativa.