Politecnico di Torino

Dipartimento di Scienze Matematiche "G. L. Lagrange"

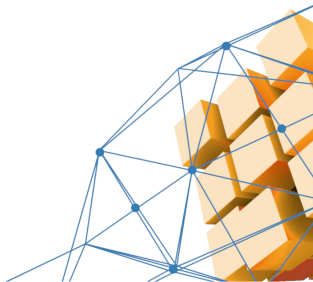
DISMA ECCELLENZA 2018•2022

Metodi di ricampionamento per l'analisi dati

Candidato:
Filippo Grobbo

Relatore:
Prof Francesco Vaccarino

Laurea in Matematica per l'Ingegneria



Introduzione al problema

Il machine learning è una branca dell'intelligenza artificiale che si occupa della costruzione di meccanismi che simulano il processo di apprendimento.
Nell'apprendimento supervisionato lo scopo principale è quello di dedurre, attraverso degli algoritmi, una funzione ϕ a partire dai dati a disposizione.

- Data set $D = \{(x_i, y_i), i = 1, 2, ..., m\}$
- x_i vettore di features, contiene le caratteristiche di una osservazione
- y_i label, rappresenta la classe di appartenenza dell'osservazione
- $\phi(x_i) \simeq y_i$

Il bilanciamento dei dati

il problema del bilanciamento dei dati riguarda la maggior parte delle applicazioni reali dove i data set presentano classi non equamente distribuite. I motivi principali per affrontare questo tipo di problemi sono:

- interesse per la classe minoritaria
- istanze rare

La combinazione di un problema sbilanciato insieme a caratteristiche intrinseche dei dati e del problema stesso porta allo sviluppo di modelli subottimali. Le principali cause di malfunzionamento degli algoritmi sono:

- small disjuncts
- mancanza di dati
- sovrapposizione delle classi
- diversa distribuzione interna delle classi

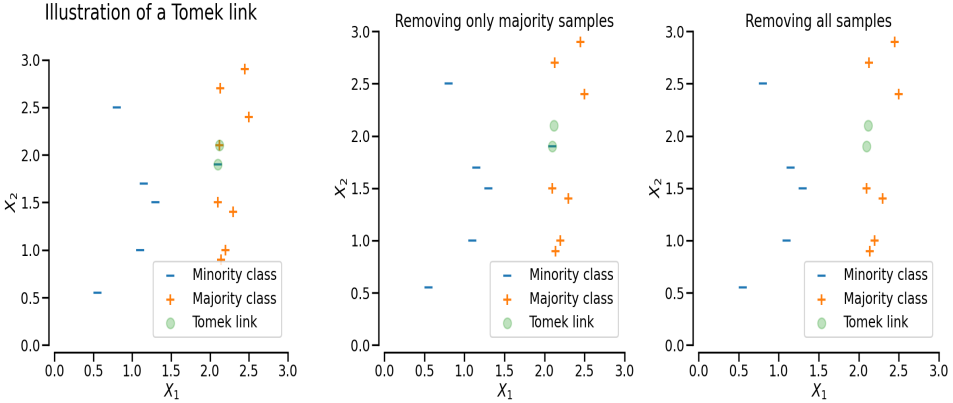
Soluzioni al problema dello sbilanciamento dei dati

- strategie di ricampionamento
  - oversampling
  - undersampling
- generazione sintetica dati
  - Smote
  - Adasyn
- assegnazione costi alle osservazioni

Oversampling & undersampling

- Random oversampling: duplica le osservazioni della classe minoritaria scegliendole in modo casuale e non singolarmente una alla volta.
- Undersampling:
  - algoritmi generativi: $D \Rightarrow D'$ tale che $|D| < |D|$ e $D' \not\subset D$
  - algoritmi selettivi: $D \Rightarrow D'$ tale che $|D| < |D|$ e $D' \subset D$.
    - tecniche di undersampling controllato: Random undersampling
    - tecniche di undersampling di pulizia: Tomek's links

Tomek's links



Limiti random oversampling & random undersampling

Random oversampling:

- overfitting
- regioni di decisione piccole e specifiche

Random undersampling:

- rimozione osservazioni importanti
- mancanza di dati

Per superare questi limiti sono state introdotte varianti delle tecniche appena presentate:

- focused sampling
- combinazione di undersampling e oversampling

Synthetic Minority Oversampling Tecnique

Algoritmo 1 Smote

funzione Smote(T,N,K)
Input T: osservazioni classe minoritaria;
N: quantità totale di oversampling;
K: parametro per la scelta delle osservazioni più vicine.
Output (N/100)*T osservazioni della classe minoritaria
if N $<$ 100 **then**
considera solo la percentuale che N rappresenta di tutto T
T=(N/100)*T
N=100
end if
N=(int)N/100 La quantità di oversampling da eseguire è un multiplo intero di 100
for $i = 1 : T$ **do**
calcola i K-nearest neighbors di i e salvali in *nnarray*
Populate(N,i,*nnarray*)
end for
end funzione

Synthetic Minority Oversampling Tecnique

Algoritmo 2 Funzione per generare osservazioni

funzione Populate(N,i,nnarray)
Input N: numero di osservazioni da creare;
i: indice dell'osservazione originale;
nnarray: array contenente le k osservazioni più vicine di i.
Output N osservazioni nuove contenute nell'array *synthetic*
Variabili *sample*: array contenente le osservazioni originali della classe minoritaria;
newindex: conteggio dati generati (inizializzato a 0);
synthetic: array contenente le osservazioni generate;
numattrs: numero di attributi presenti in ogni osservazione originale.
while $N \neq 0$ **do**
nn=random(1,k)
for $attr = 1 : numattrs$ **do**
dif=*sample*[nnarray[nn]][attr]-*sample*[i][attr]
gap=random(0,1)
synthetic[*newindex*][attr]=*sample*[i][attr] + $gap \cdot dif$
end for
newindex++
N- -
end while
end funzione

Vantaggi e estensioni di Smote

Vantaggi:

- Smote opera sullo spazio delle features
- regioni di decisione larghe e meno specifiche
- risoluzione problemi legati agli small disjuncts e alla mancanza di dati per costruzione

Successivamente sono state sviluppate delle estensioni di Smote sia per trattare data set più generici sia per risolvere le difficoltà legate alla sovrapposizione delle classi e alla diversa distribuzione interna. Tra queste vi sono ad esempio:

- Smote-NC
- Smote-N
- Borderline-Smote
- Safe-Level-Smote
- Metodi multiclasse

Adaptive Synthetic Sampling

Algoritmo 3 ADASYN

funzione Adasyn(D_{tr}, m_s, m_l)

Input D_{tr}: training set con m osservazioni $\{x_i, y_i\}$ con $x_i \in X$ e $y_i \in Y = \{-1,1\}$;

m_s: numero osservazioni della classe minoritaria;

m_l: numero osservazioni della classe maggioritaria ($m_s \leq m_l$ e $m_s + m_l = m$);

Output g nuove osservazioni della classe minoritaria ($g = \sum_{i=1}^{m_s} g_i$)

Variabili d_{th}: massimo grado tollerato di sbilanciamento;

β: parametro per specificare il grado di sbilanciamento ($\beta \in [0,1]$);

Δ_i numero di osservazioni tra i K vicini che appartengono alla classe maggioritaria

d=$\frac{m_s}{m_l}$ grado sbilanciamento

if d $< d_{th}$ **then**

G=$(m_l - m_s) \cdot \beta$ numero di osservazioni da generare per la classe minoritaria

for i=1:m_s **do**

calcola i K-nearest neighbors di x_i

r_i=$\frac{\Delta_i}{K}$, i=1,...,m_s proporzione di osservazioni maggioritarie nel vicinato di x_i

$\hat{r}_i$=$\frac{r_i}{\sum_{i=1}^{m_s} r_i}$ proporzione normalizzata

g_i=$\hat{r}_i \cdot G$ numero di osservazioni che devono essere generate per x_i

for j=1:g_i **do**

scegli in modo casuale una osservazione x_{zi} tra i K vicini di x_i

s_i=$x_i + (x_{zi} - x_i) \cdot \lambda$ nuova osservazione generata ($\lambda \in [0,1]$ numero casuale)

end for

end for

end if

end funzione

Applicazione pratica

Financial ratios per la previsione dello stato di bancarotta

Metodi utilizzati:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Support Vector Machine (SVM)
- Multi-Layer Perceptron (MLP)

Tecniche di ricampionamento utilizzate:

- Random oversampling
- Random undersampling
- Smote
- Smote+Tomek's links
- Adasyn
- Adasyn+Tomek's links

Applicazione pratica

Financial ratios per la previsione dello stato di bancarotta

Analizzando e confrontando le varie tecniche utilizzate otteniamo in un primo momento che la LDA nel caso del random oversampling risulta essere il miglior metodo predittivo. Il ruolo della PCA nella riduzione della dimensionalità inoltre non è da escludere a priori come strumento di appoggio per migliori predizioni. Un'ulteriore analisi delle features fondamentali evidenzia come gli indicatori capital structure ratio e others siano fondamentali nell'addestramento degli algoritmi. La loro rimozione infatti porta a modelli subottimali. Infine, l'utilizzo di Smote, Adasyn e Tomek's links permette un miglioramento delle prestazioni dei principali metodi utilizzati.

Grazie per l'attenzione

Politecnico di Torino