Università degli Studi di Salerno

DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE
Corso di Dottorato di ricerca
in
Economia e Politiche dei Mercati e delle Imprese
Cilco: XXXIII
Curriculum: Metodi Statistici

Sintesi della tesi (italiano)
High-Dimensional Time Series Clustering: Nonparametric Trend Estimation

Candidato:
Giuseppe Feo
Matricola 8801000030

Tutor:
Ch.mo Prof.
Francesco Giordano

Coordinatore:
Ch.ma Prof.ssa
Alessandra Amendola

Anno Accademico 2020–2021

L'era dei big data ha prodotto metodologie estese per estrarre caratteristiche/pattern da dati di serie temporali complesse. Dal punto di vista della scienza dei dati, queste metodologie sono emerse da più discipline, tra cui statistica, elaborazione/ingegneria dei segnali e informatica. Il clustering è una soluzione per classificare dati enormi quando non c'è alcuna conoscenza precedente sulle classi, ottendo così la riduzione della numerosità ad esempio.

L'obiettivo del clustering è identificare la struttura in un set di dati senza etichetta organizzando i dati in gruppi omogenei in cui la dissomiglianza all'interno del gruppo è ridotta al minimo e la dissomiglianza tra i gruppi è massimizzata. I dati sono chiamati statici se tutti i loro valori delle caratteristiche non cambiano nel tempo o se il cambiamento è trascurabile. La maggior parte delle analisi di clustering è stata eseguita su dati statici. Proprio come il clustering di dati statici, il clustering di serie temporali richiede un algoritmo o una procedura di clustering per formare cluster dato un insieme di oggetti di dati non etichettati e la scelta dell'algoritmo di clustering dipende sia dal tipo di dati disponibili che dal particolare scopo e dall'applicazione.

Considerando le serie temporali come oggetti discreti, le procedure di clustering convenzionali possono essere utilizzate per raggruppare un insieme di serie temporali individuali rispetto alla loro somiglianza in modo tale che serie temporali simili siano raggruppate nello stesso cluster. Da questa prospettiva sono state sviluppate tecniche di clustering di serie temporali, la maggior parte delle quali dipende in modo critico dalla scelta della misura della distanza (cioè della somiglianza). In generale, la letteratura definisce tre diversi approcci alle serie temporali di cluster: (i) *Shape-based clustering*, il clustering viene eseguito in base alla somiglianza delle forme, in cui le forme di due serie temporali sono abbinate utilizzando contrazioni e decontrazioni non lineari degli assi temporali; (ii) *Feature-based clustering*, le serie temporali grezze vengono trasformate nel vettore di caratteristiche di dimensione inferiore dove, per ogni serie temporale, viene creato un vettore di caratteristiche di lunghezza fissa e uguale (di solito un insieme di caratteristiche statistiche); (iii) *Model-based clustering* assume un modello matematico per ciascun cluster e tenta di adattare i dati al modello assunto.

La scelta di un metodo di rappresentazione appropriato può essere considerata la componente chiave che influisce sull'efficienza e sull'accuratezza della soluzione di clustering. L'elevata dimensionalità e il rumore sono caratteristiche della maggior parte delle serie temporali, di conseguenza, i metodi di riduzione della dimensionalità vengono utilizzati nel clustering delle serie

temporali per affrontare questi problemi e promuovere le prestazioni. La composizione del trend delle serie temporali è un argomento molto importante nell'analisi dei dati, specialmente nella letteratura più recente sul clustering delle serie temporali ad alta dimensionalità. Il controllo della composizione del trend è il primo passo per un'ulteriore analisi statistica condotta su una serie temporale. Infatti, molte delle procedure di clustering proposte in letteratura si basano sul presupposto che tutte le serie storiche considerate seguano la stessa struttura di trend. Quest'ultimo può essere assente, lineare o non lineare. In realtà, la vera struttura del trend è sconosciuta, quindi è necessaria una procedura che permetta questa distinzione prima di qualsiasi analisi di clustering. Con questo in mente, la tesi proposta mira a colmare questa lacuna.

In particolare, la proposta discussa in questa tesi riguarda un'analisi embrionale per effettuare una corretta ulteriore analisi di clustering su serie temporali. Precisamente, riguarda la classificazione di serie storiche non stazionarie, dove la non stazionarietà è data dalla presenza di un trend deterministico, guardando la derivata prima del trend in un contesto di alta dimensionalità e senza richiedere una forma prestabilita per il trend. Ciò si ottiene mediante uno stimatore non parametrico che ha una forma molto semplice. L'idea è quella di classificare le serie temporali controllando la derivata prima del trend. Se il trend è costante, allora la sua prima derivata è zero, se esso è lineare, allora la sua prima derivata è costante. Se non si verifica nessuno dei precedenti, il trend è ovviamente non lineare e quindi la sua derivata prima non sarà costante. In questo modo le serie temporali possono essere suddivise in tre gruppi. Questo approccio può essere incluso nella categoria del "raggruppamento di serie temporali Feature-based", poiché la composizione del trend può essere considerata una caratteristica della serie storica. Una volta classificate le serie storiche sarà possibile applicare la tecnica di clustering più appropriata.

Supponiamo di osservare p (il quale può tendere ad infinito come funzione dell'orizzonte temporale) serie storiche indipendenti della forma

$$Y_{it} = m_i(t/T) + \varepsilon_{it}, \quad i = 1, \ldots, p; t = 1, \ldots, T \tag{1}$$

dove $m_i : [0, 1] \to \mathbb{R}$ sono funzioni di trend sconosciute e $\{\varepsilon_{it}\}_{t=1}^{T}$ processi strongly mixing a media zero. Al fine di partizionare tali serie temporali in base alla loro composizione del trend (costante, lineare o non lineare), si

può stimare la derivata prima del trend utilizzando uno stimatore non parametrico in una delle condizioni di dipendenza meno restrittive del termine di errore. Lo stimatore non parametrico proposto per la derivata prima del trend, al punto $x \in [0, 1]$, ha la forma

$$\hat{\beta}(x) = \frac{1}{Th^2} \sum_{t=1}^{T} K_h(t/T - x)(t/T - x)Y_t, \tag{2}$$

dove $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$ con $K(\cdot)$ una funzione kernel simmetrica e Lipshitz continua con supporto limitato, $h = h_T > 0$ è il bandwidth tale che $Th^4 \to \infty$ per $T \to \infty$. Lo stimatore proposto, basato sulla linea guida dello stimatore Local Polynomial con disegno fisso, ha la caratteristica interessante di essere proporzionale alla derivata prima reale poiché il suo bias dipende solo da una quantità nota per $T \to \infty$.

Sotto la ragionevole ipotesi che il numero di serie storiche con andamento non lineare sia finito, la procedura di partizione proposta consiste in due fasi. Nella prima, lo stimatore proposto viene testato per essere zero o meno, il che consente di distinguere le serie storiche con andamento costante. Nella seconda, la differenza tra lo stimatore in punti diversi viene utilizzata in un approccio di screening per effettuare l'ulteriore partizione lineare/non lineare delle restanti serie temporali della fase precedente. In altre parole, la prima fase viene utilizzata per selezionare le serie storiche con andamento costante mediante una procedura di verifica mentre la seconda è una procedura di screening che fornisce l'insieme che contiene, con probabilità tendente a 1, l'insieme vero delle serie storiche con trend non lineare. L'algoritmo di seguito fornisce i dettagli dei vari passaggi e mostra la facile implementazione dell'intera procedura.

Sono stati condotti studi simulativi con impostazioni diverse per il termine di errore, T e h per verificare non solo ogni parte della procedura singolarmente, ma anche l'intera procedura. I risultati ottenuti confermano quanto teoricamente dimostrato per la procedura in due fasi.

Infine, è stato proposto un esempio di applicazione su dati reali ("Smart meter data from London" disponibile su `https://www.kaggle.com`) per mostrare l'effettiva bontà e necessità della procedura prima di applicare una cluster analisi su serie temporali.

L'utilizzo dell'approccio citato presenta molteplici vantaggi: (i) dal punto di vista matematico, è abbastanza intuitivo l'uso della derivata prima per

Algorithm Classify HD Time Series by Trend

Set $U := \{1, \ldots, p\}$, $C_1 = C_2 = C_3 = \emptyset$
Set the parameters α, s and h_i, $i \in U$
for $i \in U$ **do**
    Perform the "Trend/NoTrend Test Statistic" $\hat{I}_{\beta,i}$
    if $\hat{I}_{\beta,i} < \chi^2_{(1-\alpha/p,k_T)}$ **then**
        Set $C_1 := C_1 \cup \{i\}$
Set $U := U \backslash C_1$
if $U = \emptyset$ **then**
    return C_1, C_2, C_3
else Perform the "Lin/NoLin Statistic" $\hat{I}_{D,i}$, $i \in U$, and sort them as $\hat{I}_{D,\sigma(1)} \geq \cdots \geq \hat{I}_{D,\sigma(p_2)}$
    Set $C_3 := \{\sigma(1), \ldots, \sigma(s)\}$ and $C_2 := U \backslash C_3$
    return C_1, C_2, C_3

evidenziare la linearità di una funzione; (ii) si può affermare se un trend è lineare o meno senza imporre un modello matematico predefinito; (iii) questo tipo di procedura effettua una partizione dell'insieme delle serie temporali date che può essere utilizzata in un'ulteriore analisi come punto di partenza (ovvero fornisce una conoscenza preliminare utile sulla composizione del trend per un'analisi di clustering più approfondita); (iv) non impone restrizioni alla composizione del trend come quelle che si impongono quando si verifica la presenza del parallelismo; (v) dà garanzie matematiche nell'ambito ad alta dimensionalità poiché è consistente nel caso in cui il numero di serie storiche tenda all'infinito, cioè $p = o\left(T^{1/2}/\log T\right)$.

Futuri sviluppi della procedura proposta riguardano la trasformazione della seconda fase in una procedura di selezione che consenta di identificare con maggiore precisione il vero insieme di serie storiche con andamento non lineare e l'aumento della dimensionalità ottenibile raggiunta dalla procedura.