

1. Introduction

"Clustering is a fundamental technique of unsupervised learning where a set of data points has to be divided into homogenous groups of units which exhibit a similar behaviour relative to a target variable".

"homogenous" non è il termine corretto, dovrebbe essere "homogeneous". Inoltre è una frase che introduce molti termini diversi fra loro (data points, units, target variable), causando un po' di confusione. Utilizzerei un solo termine (e.g., data points), dicendo che nel clustering si tenta di avere data points simili nello stesso cluster e di avere che data points di cluster diversi non sono simili. Potresti trovare qualche chiacchiera di esempio da cui prendere spunto in "An Introduction to Statistical Learning" di Hastie e Tibshirani, nel capitolo dedicato ai metodi di clustering.

"Thus, the development of these models finds its natural context within the Bayesian framework."

UP 1

Questa frase mi sembra che crei un collegamento logico un po' troppo forte tra metodi model-based e il framework Bayesiano. Infatti, anche in un framework frequentista si può definire tranquillamente un modello per fare clustering.

Mi muoverei su una frase un po' più "soft", dicendo che il framework Bayesiano permette di sviluppare modelli probabilistici in grado di fornire una stima del clustering e della sua incertezza.

The effectiveness of Bayesian clustering models is particularly evident in the complex scenarios of spatio-temporal data, in which observations are collected over time and across various spatial locations. This scenario is inherently complex due to the interplay between spatial and temporal dimensions, a complexity that is further compounded when covariates are also available. Nevertheless, Bayesian models effectively integrate all these levels of information into the model, thereby improving the quality of the results.

Non mischierei il discorso sul fatto di usare metodi Bayesiani con il tipo di dati che stiamo utilizzando. In questo paragrafo, direi che stiamo lavorando con serie temporali caratterizzate da una localizzazione spaziale e da covariate, che sono anch'esse serie temporali. Introdurrei anche il fatto che si utilizzerà un dataset sul PM10, così da dare subito un esempio al lettore per aiutarlo a capire di cosa si sta parlando.

In this work, we tackled...

"We tackle...": usiamo sempre la forma verbale presente.

Introdurrei una frase per dire che l'implementazione più efficiente è necessaria perché in questo contesto si hanno molte informazioni su cui lavorare. In questo modo "giustificiamo" il perché hai lavorato su questo progetto e la sua utilità.

UP 2

2. Description of the model

The main clustering tool employed in Bayesian clustering methods is a mixture model, which can naturally be extended to an infinite-mixture model within Bayesian nonparametrics.

Aggiungerei una frase per sottolineare il fatto che la scelta di andare verso una infinite mixture model (e quindi di utilizzare un modello nonparametrico) evita che ci sia bisogno di selezionare o stimare a priori il numero di cluster. Puoi trovare qualche chiacchiera a riguardo in Grazian (2023), A review on Bayesian model-based clustering.

In the spatio-temporal context, where we analyse the sequence of partitions ρ_1, \dots, ρ_T , i.e. the clusters of the n experimental units at each time instant $t = 1, \dots, T$, this choice is not always effective

Questa cosa viene introdotta e data per scontata, ma potrebbe non esserla per un lettore. Quindi, la introdurrei per bene a parole tentando di non utilizzare la notazione, che viene spiegata solo successivamente.

In generale, vedo che nel primo paragrafo della sezione 2 (prima della sezione 2.1) stai seguendo il discorso dell'introduzione di Page et al. (2022). Può andare bene secondo me, ma sistemerei un paio di cose:

- Non introdurrei il concetto di random probability measures, visto che poi non se ne parlerà più per il resto dell'articolo. Piuttosto, direi soltanto che nei modelli tradizionali *inducono* il rapporto tra le partizioni, ma non lo modellano direttamente. Invece, il DRPM lo modella direttamente, creando un maggiore rapporto temporale che può essere controllato tramite un parametro (con riferimento a Figure 2 di Page et al. (2022)).
- In fact, in infinite mixture models, partitions are only induced by the random partition model.

Questa frase non è corretta, perché anche il DRPM può essere visto come un infinite mixture model marginalmente (vedi Proposition 1 di Page et al. (2022)). Piuttosto, direi che la maggior parte dei modelli temporali di BNP inducono la dipendenza temporale in maniera indiretta, mentre il DRPM la induce in maniera diretta.

Here, $P(\rho_1)$ is an exchangeable partition probability function (EPPF), implemented by (Page et al., 2022) through a Chinese restaurant process in the form

Inserirei nella notazione il fatto che $|S_{jt}|$ indica la cardinalità del cluster, inoltre nell'equazione (2) credo che si debba scrivere $|S_{j1}|$.

A differenza di quanto si fa in Page et al., io rappresenterei l'equazione (2) in forma generale (vedi equazione (1) di Page and Quintana (2016), Spatial Product Partition Model), dicendo poi che si uno dei casi più famosi è quello del Chinese Restaurant Process, cioè dell'EPPF del Dirichlet Process.

In equazione (3) manca una parentesi chiusa.

Darei un minimo di spiegazione testuale riguardo il modello dell'equazione (4) (e (5)). In particolare, sottolineerei il perché della scelta di modellare η_{1i} con questa trasformazione di una Laplace (i.e., troncatura tra -1 e 1 e code abbastanza poco pesanti, così da non dare tutto il legame temporale alla likelihood), così come il fatto che i θ_t servono a dare un ulteriore legame temporale tra i valori unici dei cluster. Senza un minimo di spiegazione testuale temo che un modello del genere sia difficile da leggere per un lettore che non sia molto legato a questi modelli.

Each parameter will either require a Gibbs step or a Metropolis step, depending on the conjugacy of the prior associated with that parameter.

Non è corretto, perché il Gibbs step consiste nel fare sampling dalla prior, cosa che avviene sia nel caso di coniugatezza che nel caso di Metropolis. Direi piuttosto che il sampling di ogni altro parametro dalla sua full conditional in alcuni casi è coniugato, mentre in altri richiede uno step di Metropolis.

Aggiungerei anche una frase per dire che la derivazione di queste full conditional può essere fatta con calcoli standard e per rimandare il lettore interessato alla corrispondente sezione della tesi, dove può trovarli in forma estesa.

Fornirei qualche informazione in più (non tutti i dettagli) riguardo l'update di γ e di ρ , un po' come viene fatto in Page et al. (2022). Visto che è una parte abbastanza importante dell'algoritmo, credo che il lettore debba capire cosa è necessario fare.

Nella sezione 2.3, utilizzerei la notazione s_j^* , visto che sappiamo che le coordinate non cambiano mai nel tempo.

la t era per riferirsi al cluster $S_{\{jt\}}$, perché i cluster sono identificati dalla coppia (j,t) usando la notazione precisa spazio-temporale
Cioè è vero che "le coordinate non cambiano nel tempo" ma cambiano le coordinate del cluster j al tempo t

In the JDRPM code, we implemented

⇒ "we implement"

is the covariate similarity, parametrised by a set of parameters \mathcal{C} , which measures the similarity of the covariates values s_{jt}^* .

C'è un typo, le covariate values dovrebbero essere X_{jtr}^* .

3. Implementation

As a reference, our development process comported an improvement by 74% in execution time and by 88% in the memory allocated when comparing an early-stage version to the final version of the Julia code.

Mi sembra una frase un po' a doppio taglio: da un lato può mostrare che ci sono stati dei miglioramenti, ma dall'altro un lettore potrebbe pensare che il codice iniziale fosse scritto male. La toglierei: il fatto che hai fatto dei miglioramenti studiati mi sembra che sia già evidente dal resto della sezione, quindi questa frase rischia solo di metterti nei guai.

4. Analysis of the models

Inserirei l'immagine del "Kriging" con le tre time series che vengono predette: mi sembra una bella immagine per far capire cosa succede e puoi parlare più facilmente del punto che viene predetto male.

Regarding the kriging analysis,

Non utilizzerei il termine Kriging in questa frase e nel resto del documento. Infatti, Kriging è un metodo ben preciso (tra l'altro, frequentista) che non c'entra molto con quello che facciamo noi, se non per il fatto che si può fare predizione in una nuova location. Quindi, è un termine che utilizziamo tra noi per capirci, ma che in realtà non sarebbe corretto.

Inserirei anche l'immagine dei missing points con i 95% credible intervals, anche quella mi sembra un bel colpo d'occhio per mostrare che il modello funziona.

wind speed, precipitation levels, and boundary layer height.

Aggiungerei una citazione per giustificare la scelta di queste covariate e una breve spiegazione di cosa sia il boundary layer height.

Secondo me si può cambiare la scala verticale dei boxplot della figura 3, così da vedere meglio le differenze tra i boxplot. Per esempio, in WE_tot_precipitation, la scala verticale può essere tra -1 e 1.5.

Although incorporating multiple covariates introduces the possibility of divergent "clustering suggestions,"

Questo è un problema che va spiegato meglio, altrimenti così non è di facile comprensione per chi legge. Direi che l'aggiunta di covariate nel clustering spesso comporta una maggiore frammentazione in cluster dell'intero dataset, dovuta al fatto che ci sono più fattori in gioco che possono rendere due unità dissimili fra loro. Per esempio, in Figura 3 si può notare come i cluster 3 e 4 abbiano valori simili per PM10 ma completamente diversi per quanto riguarda le precipitazioni.

La stessa cosa, come hai scritto giustamente tu, si può vedere nelle figure 4 e 5, in cui JDRPM identifica più cluster rispetto a CDRPM, "spezzando" alcuni cluster di quest'ultimo.