Modellare la complessità di serie temporali biomedicali con grafi e metodi di integrazione

Veronica Tozzo, Federico Tomasi, Alessandro Verri, Annalisa Barla
DIBRIS - Università degli Studi di Genova
{veronica.tozzo,federico.tomasi}@dibris.unige.it, {alessandro.verri, annalisa.barla}@unige.it

Abstract

L'analisi di dati biomedicali caratterizzati da una componente longitudinale richiede l'identificazione di *pattern* di interazioni tra variabili o di *biomarker* che siano in grado di rappresentare il segnale in maniera compatta. Per questo motivo é importante usare modelli di learning che considerino la dinamicità del dato e abbiano la possibilità di integrare misure eterogenee. Ad esempio possiamo considerare i modelli grafici per studiare l'evolvere dei dati nel tempo anche integrando le reti ottenute per ottenere una piu' robusta stima del comportamento delle variabili in azione. Metodi di integrazione sono anche adatti per l'identificazione di biomarker specifici di un segnale fisiologico o patologico.

1 Introduzione

Negli ultimi anni le tecniche di acquisizione e archiviazione di dati hanno consentito una rapida crescita nell'ammontare di dati che possono essere analizzati. Questo è avvenuto in molteplici aree applicative, tra cui finanza, sociologia, genomica e biomedicale. Specialmente nelle ultime due aree è sempre di maggior interesse cercare di studiare l'andamento del dato nel tempo per cercare di identificare dei pattern identificativi di un certo stato. Questi pattern possono essere a livello di interazione di variabili che che entrano in gioco tanto quanto l'identificazione di biomarker, temporali e non, che siano in grado di descrivere il segnale nella sua totalità.

Al fine di analizzare questi dati è necessario sviluppare modelli di learning che siano in grado di abbracciarne la complessità e districarne il funzionamento.

Un esempio di modello complesso coinvolge la rappresentazione del comportamento delle variabili in specifiche condizioni biologiche, farmacologiche o ambientali e la loro variabilità nel tempo.

A questo fine diverse tecniche possono essere utilizzate, ad esempio se si vuole ottenere una network di interazione tra le variabili che agiscono in un particolare processo è possibile ricorrere a metodi di inferenza di grafi che modellano le interazioni tra variabili (nodi) come archi in una rete.

La modellazione grafica offre una rappresentazione compatta ed efficiente che semplifica l'analisi del sistema, specialmente se effettuata al variare del tempo.

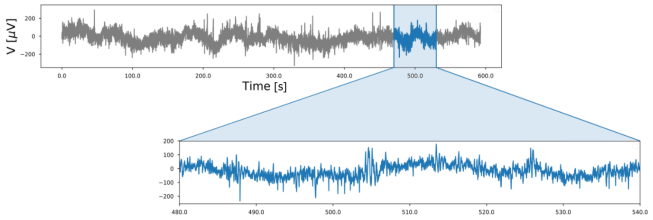

Figura 1: Esempio di segnale temporale per una variabile in pazienti affetti da epilessia focale.

Spesso nei modelli grafici si assume che le variabili abbiano la stessa natura. Negli ultimi anni la ricerca si è mossa verso lo studio di sistemi con variabili di tipo etereogeneo e quindi verso l'integrazione di reti.

Ad esempio, le interazioni in un sistema dinamico possono essere modellate a diverse frequenze (per esempio usando una trasformata wavelet) in modo da studiare l'influenza della variazione di frequenza. Questo ci permetterebbe di identificare frequenze più significative per una determinata condizione biologica (o medica) e procedere quindi a caratterizzare il dato con queste frequenze.

In quello che segue presentiamo metodi che possono essere integrati per fornire una rappresentazione di un sistema dinamico complesso. Questi metodi sono particolarmente indicati per l'utilizzo in ambito biomedicale grazie alla possibilità di rappresentare un alto numero di variabili in presenza di un ristretto numero di campioni.

2 Dati e metodi

Il tipo di dato considerato nelle analisi è una collezione di serie temporali su variabili che sono legate all'interno di un sistema complesso. Ad esempio in Figure 1 è rappresentato la variazione dell'impulso elettrico in un determinato punto del cervello (canale) di un paziente affetto da epilessia. Date le misurazioni di canali posizionati in diverse regioni del cervello lo scopo vuole essere identificare le connessioni funzionali di queste regioni. Ad esempio, in Figure 2, dati tre canali vogliamo capire come la loro correlazione varia nel tempo.

2.1 Inferenza di grafi

Per identificare le relazioni di variabili nel tempo si possono utilizzare diversi metodi regolarizzati basati sul *graphi-*

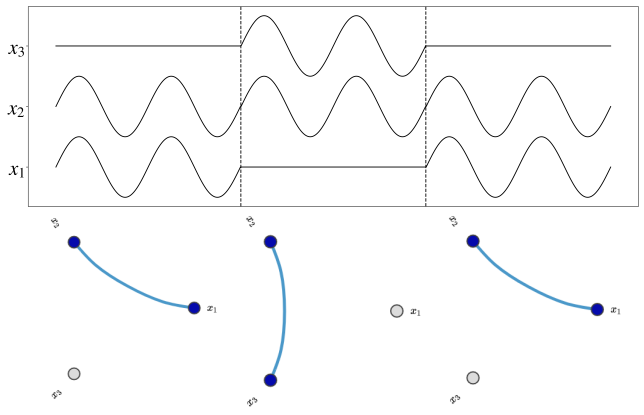

Figura 2: Esempio di rappresentazione grafica dinamica per tre segnali temporali correlati in maniera variabile.

cal lasso che assumono un comportamento consistente della rete in punti consecutivi nel tempo [Friedman *et al.*, 2008; Hallac *et al.*, 2017]. Questi metodi, possono essere estesi per considerare la presenza di variabili latenti che influenzano la misurazione del segnale. Infatti, in sistemi complessi, come quelli biologici, la presenza di variabili che non possono essere misurate ma che agiscono nel sistema è frequente [Tomasi *et al.*, 2018]. Questi metodi si sono dimostrati efficaci per l'analisi e la comprensione di serie temporali e possono ulteriormente essere estesi per considerare segnali periodi e reti a due layer [Tozzo *et al.*, 2018].

2.2 Metodi di integrazione

L'identificazione di biomarker che siano rappresentativi per uno stato patologico può essere guidata da prior knowledge sul problema oppure essere completamente data-driven. In questo secondo caso, per quanto riguarda i segnali temporali è possibile cercare di identificare le frequenze più rilevanti all'interno della patologia o condizione biologica. Per esempio, nel contesto di pazienti affetti da epilessia, è possibile applicare un'estensione di multiple kernel learning [D'Amario *et al.*, 2018] per selezionare le frequenze più utilizzate nella classificazione di canali (epilettico - non epilettico). Le frequenze identificate possono poi essere selezionate per inferire interconnessioni tra i canali a diversi livelli di frequenza, questo permette di studiare se i grafi ottenuti includono informazione specifica sullo stato patologico dettata da quella frequenza in quanto, nel cervello, esistono diverse bande di frequenza che sono legate a diversi stati di veglia/sonno/comportamento. È inoltre possibile, se si vuole ottenere una panoramica più generale integrare i risultanti grafi in uno solo utilizzando tecniche di integrazione basate su Non-negative Matrix Factorization [Kuang *et al.*, 2012].

3 Implementazione

L'implementazione dei modelli grafici che evolvono nel tempo è disponibile in REGAIN, una libreria Python disponibile sotto BSD-3-Clause all'indirizzo https://github.com/fdtomasi/regain. Allo stesso modo la pipeline per l'analisi di dati di epilessia focale è disponibile all'indirizzo https://github.com/fdtomasi/multikernel.

4 Conclusioni

Nell'ottica di estrapolare informazioni da dati multivariati e dinamici l'utilizzo di metodi in grado di raggiungere diversi livelli di astrazione è fondamentale per identificare interconnessioni e cause che possono portare a condizioni di non normalità. In questo contesto i modelli grafici, in particolare quelli regolarizzati, sono indicati per la modellazione di legami di correlazione e anche causali e pertanto, insieme a metodi in grado di integrare diverse fonti di informazione, devono essere ulteriormente studiati e migliorati. Sono anche indicati per gestire i problemi legati alla scalabilità computazionale e gestire un numero di variabili sempre maggiore.

Riferimenti bibliografici

[D'Amario *et al.*, 2018] Vanessa D'Amario, Federico Tomasi, Veronica Tozzo, Gabriele Arnulfo, Annalisa Barla, e Lino Nobili. Multi-task multiple kernel learning reveals relevant frequency bands for critical areas localization in focal epilepsy. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, e Jenna Wiens, editors, *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 348–382, Palo Alto, California, 17–18 Aug 2018. PMLR.

[Friedman *et al.*, 2008] Jerome Friedman, Trevor Hastie, e Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[Hallac *et al.*, 2017] David Hallac, Youngsuk Park, Stephen Boyd, e Jure Leskovec. Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 205–213, New York, NY, USA, 2017. ACM.

[Kuang *et al.*, 2012] Da Kuang, Chris Ding, e Haesun Park. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM international conference on data mining*, pages 106–117. SIAM, 2012.

[Tomasi *et al.*, 2018] Federico Tomasi, Veronica Tozzo, Saverio Salzo, e Alessandro Verri. Latent Variable Time-varying Network Inference. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '18, New York, NY, USA, 2018. ACM.

[Tozzo *et al.*, 2018] Veronica Tozzo, Federico Tomasi, Margherita Squillario, e Annalisa Barla. Group induced graphical lasso allows for discovery of molecular pathways-pathways interactions. *Machine Learning for Health (ML4H) Workshop at NeurIPS 2018*, November 2018.