Visual Analytics 2020 - Report

Cinquini Martina, matricola 522482

Settembre 2020

Abstract

Il progetto nasce con l'obiettivo di sviluppare un'interfaccia grafica in grado di fornire all'utente un sistema di supporto visuale e interattivo per comprendere il motivo per cui le variabili osservate sono identificate in una determinata relazione di causa-effetto dall'algoritmo di discovery. L'applicazione è strutturata in tre sezioni: la prima contestualizza l'algoritmo; la seconda esplora nel dettaglio la procedura applicandola a dataset bivariati; la terza permette di investigarne la generalizzazione utilizzando un dataset multivariato.

1 Introduzione

La scoperta dei rapporti di causalità, che regolano i fenomeni osservati, è una necessità connaturata dell'indole umana. Al giorno d'oggi, la grande quantità di dati disponibili può aiutare a soddisfare questa innata curiosità. L'identificazione delle strutture causali è quindi diventata un interessante campo di ricerca in molti settori tra cui la medicina, l'epidemiologia e l'economia. Conoscere la causalità, infatti, aiuta a interpretare i dati, a formulare e a testare ipotesi e a spiegare le teorie di modellazione. In relazione all'ultimo aspetto, la possibilità di questa conoscenza, sta sviluppando una crescente attenzione anche nell'ambito dell'Explainable Artificial Intelligence che mira a costruire algoritmi interpretabili e trasparenti.
Dedurre questi meccanismi risulta tuttavia impegnativo perché le relazioni causali non possono essere verificate esclusivamente sulla base di modelli matematici, in quanto informazioni supplementari sono necessarie per ridurre l'ambiguità generata da fattori non considerabili da test statistici.
Tra le modalità per individuare le relazioni, vi è quella di condurre una sperimentazione controllata che assegni casualmente i partecipanti in un gruppo di trattamento o in un gruppo di controllo. Tuttavia, questi esperimenti sono lunghi e molto costosi. Inoltre, possono coinvolgere solo un determinato numero di soggetti che potrebbero non essere sufficientemente rappresentativi della popolazione di interesse. Infine è necessario considerare anche le questioni etiche di queste tipologie di studi che ne limitano ampiamente le applicazioni.
Al fine di evitare questi problemi, è auspicabile lo sviluppo di metodologie che si basino puramente sui dati osservazionali. Negli ultimi anni, sono stati proposti vari algoritmi di discovery in grado di stimare l'effetto causale in base a determinati presupposti. Se i i dati sono a valore continuo, vengono comunemente applicati metodi basati su modelli causali che ipotizzano che gli effetti siano funzioni lineari delle loro cause più un rumore gaussiano indipendente [4]. Sebbene i presupposti della linearità e della gaussianità siano matematicamente convenienti, non sempre sono realistici.
Il progetto si incentra su una procedura specifica proposta in origine da Hoyer et el. [3] che dimostra che il framework lineare può essere generalizzato a modelli non lineari. In particolare, le *non linearità* nel processo di generazione dei dati consentono di identificare informazioni più accurate del sistema causale. L'interfaccia implementata fornisce una visualizzazione interattiva che permette di esplorare passo dopo passo l'algoritmo di discovery. L'obiettivo è la costruzione di una profonda comprensione del motivo per cui la procedura identifica proprio quel determinato modello causale tra i diversi scenari possibili.

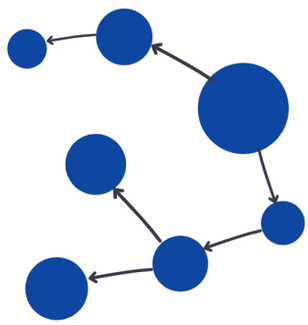

Figure 1: Grafo diretto aciclico rappresentante un modello causale

2 Analisi visuale delle relazioni causali non lineari

Nel campo della Visual Analytics, si sta gradualmente sviluppando un interesse nell'analisi visuale della causalità. Di particolare rilievo, è infatti la maggiore acquisizione di conoscenza dell'utente nel processo di identificazione delle relazioni causali grazie alla possibilità di esplorazione e di interazione. A tal proposito sono stati progettati vari approcci visivi [2, 5, 1, 6], tuttavia le relazioni causali non lineari sono state raramente studiate direttamente. Con l'intento di considerare le precedenti visualizzazioni come punto di partenza, il progetto cerca di fornire all'utente un insieme di strumenti interattivi per esaminare anche le relazioni causali non lineari.
Nella fase di progettazione dell'applicazione, si è pensato di creare una struttura che ne permettesse l'utilizzo sia ad utenti esperti del dominio di interesse sia a coloro che si approcciano al mondo dell'identificazione delle causalità senza nessuna conoscenza pregressa. Al fine di soddisfare entrambe le esigenze, si è ritenuto opportuno organizzare in sezioni distinte i contenuti principali dell'intefaccia. Nello specifico, la contestualizzazione dell'algoritmo *"Nonlinear causal discovery with additive noise methods"* nell'ambito della causal discovery e la visualizzazione della procedura prima applicata al caso bivariato e successivamente al caso multivariato. Questi contenuti sono raggruppati in modo da rendere chiaramente distinguibili le informazioni principali da quelle secondarie o di supporto, ponendo quelle rilevanti immediatamente visibili. La suddivisione, in base al numero di variabili del dataset, è dovuta alla scelta di voler realizzare un'interfaccia in grado di visualizzare più fedelmente possibile l'approccio della procedura di causal discovery descritto nel paper [3].
Per realizzare gli elementi interattivi presenti nel progetto, i dati sono stati generati in base all'implementazione in Python dell'algoritmo applicata ad una varietà di dataset simulati e reali. In particolare, per i dati reali sono stati scelti quattro dataset di diversi domini, ognuno costituito da campioni di una coppia di variabili casuali dipendenti in cui una variabile è nota per causare l'altra, mentre per quelli sintetici si è optato per riprodurre il meccanismo generativo già sperimentato [3] che produce un dataset con quattro variabili.
Infine oltre alla comprensione della procedura, quest'interfaccia permette di: riconoscere le simmetrie dei dati osservazionali; mostrare la capacità del metodo di trovare il modello corretto quando tutte le ipotesi sono soddisfatte; e di visualizzare il grafo causale dedotto risultante.

3 Implementazione interfaccia

In questa sezione sono illustrate le tecniche utilizzate per l'implementazione di widget visuali e interattivi e l'analisi di ciascuno di questi.

3.1 Tecniche utilizzate

Come ambiente di sviluppo, si è utilizzato Node.js, un framework runtime in Javascript che permette di realizzare applicazioni di rete scalabili e NPM, un gestore di pacchetti che permette di includere, rimuovere e aggiornare le librerie.
L'interfaccia utente è stata realizzata grazie all'uso di Vue.js, un framework javascript di tipo progressivo che si basa sul rendering dichiarativo e sulla composizione dei componenti.
Inoltre, si è scelto di integrare la libreria Vue Router, componente ufficiale di Vue.js per il routing e la navigazione delle pagine. Esso offre molteplici funzionalità tra cui la mappatura nidificata delle view, la configurazione del router modulare basata su componenti, e la visualizzazione degli effetti di transizione.
Per sfruttare queste caratteristiche, il progetto è strutturato con una parent route, App.vue, che indica dove eseguire il rendering delle tre componenti figlie che sono rispettivamente home.vue, bivariateCase.vue e multivariateCase.vue. In particolare, le ultime due vengono utilizzate per la manipolazione dei dati da passare alle componenti annidate in cui sono implementati i grafici. Riguardo all'esplorazione interattiva dei dati, si è scelto di utilizzare le librerie Javascript Plotly.js e D3.js.
Infine particolare attenzione è data alla realizzazione grafica dell'applicativo, pensata per adattarsi agli standard Web di usabilità e accessibilità, attraverso una struttura chiara e logica. Questa implementazione consente ai contenuti di essere significativi e facilmente identificabili e garantisce una coerenza visuale sia all'interno della pagina navigata sia tra le diverse pagine dell'applicazione. Per lo sviluppo si è utilizzato Material Design Bootstrap, an UI kit integrato in Vue.js.

3.2 Widget visuali e interattivi

La prima visualizzazione dell'applicazione, collocata nella sezione del caso bivariato, è uno scatter plot (figura 2) realizzato con la libreria D3.js. Si è scelto di utilizzare questo grafico per dare la possibilità all'utente di esaminare i valori di una variabile, tracciati in funzione di un'altra. In particolare, gli è richiesto di provare ad identificare quale variabile è la causa e quale l'effetto, analizzando solamente il grafico senza avere nessuna conoscenza pregressa riguardo ai dati. La risposta al quesito è mostrata tramite un pulsante interattivo posto alla desta dello scatter plot. Lo scopo di questa visualizzazione è quindi fornire all'utente una visione d'insieme del problema prima di considerare nello specifico l'algoritmo di discovery.

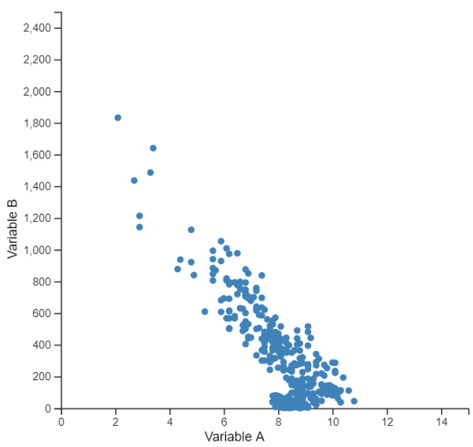

Figure 2: Scatter plot rapprentante la relazione causale tra due variabili

La visualizzazione successiva è un grafo diretto aciclico (figura 3) costituito da due nodi, ossia le variabili osservate, e da un arco che li collega. La presenza dell'arco implica una relazione causale mentre la sua direzione identifica l'effetto dalla causa. Questo layout 2D interattivo, implementato con D3.js, rappresenta la struttura causale dedotta dai dati osservazionali. Nel caso riportato in figura 3, riguardante il dataset Abalone, possiamo concludere che l'età, ossia gli anelli del mollusco, causano la sua lunghezza.

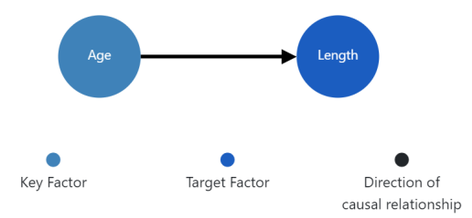

Figure 3: Grafo diretto aciclico rappresentante la struttura causale stimata

Per comprendere il motivo per cui, la procedura identifica proprio quella determinata struttura causale tra i diversi scenari possibili, è utile poter esaminare anche i risultati della regressione e i relativi residui di entrambi i modelli direzionali. A tal fine, si sono implementati con Plotly.js quattro grafici di dispersione, rispettivamente due per il forward fit (figura 4) e due per il backward fit. Negli scatter plot che mostrano la regressione, è possibile esplorare le osservazioni, le predizioni e l'intervallo di confidenza al 95%.

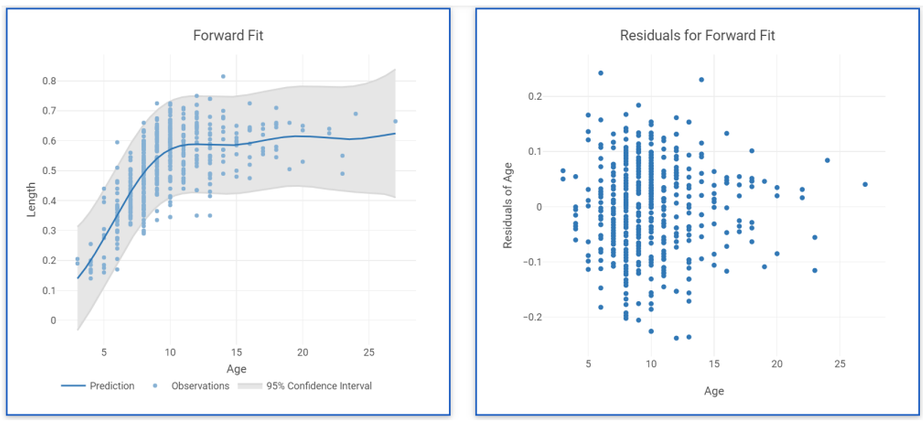

Figure 4: Dataset Abalone: (a) forward fit corrispondente a "età (anelli) causa lunghezza"; (b) residui del forward fit;

Le successive visualizzazioni riguardano la procedura generalizzata ad un numero N di variabili. In tal caso, un modo per stimare il modello causale è quello di utilizzare una ricerca esaustiva. Tuttavia, la complessità della procedura aumenta in modo esponenziale in base al numero di variabili. Tenendo conto di ciò, il metodo ANM è fattibile solo per reti molto piccole.
Si è scelto di rappresentare visualmente questo aspetto per conferirgli la giusta importanza, catturando l'attenzione dell'utente con una visualizzazione interattiva. Questo tipo di grafico, implementato con D3.js, è pensato per dare un'idea immediata della crescita esponenziale delle combinazioni e può essere esplorato utilizzando lo slider collocato nella parte superiore.

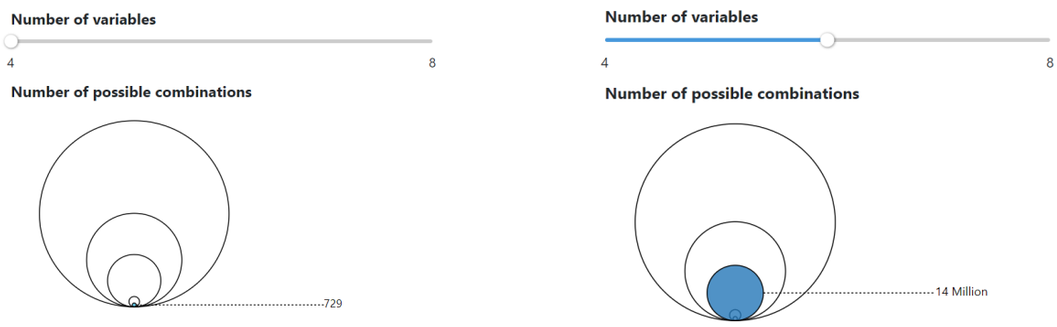

Figure 5: Rappresentazione della ricerca esponenziale del numero di combinazioni all'aumentare del numero di variabili da considerare; (a) quattro variabili (b) sette variabili

L'ultima visualizzazione concerne l'esplorazione dinamica dei risultati degli esperimenti sui dati sintetici. Si è scelto di realizzare, con Plotly.js, una heatmap che permette di visualizzare i dati attraverso le variazioni di colorazione. Per poter comprendere queste ultime in modo appropriato, al lato destro è stata collocata una legenda. Le righe e le colonne indicano le variabili del dataset, mentre i dati contenuti all'interno delle celle si basano sulla relazione tra le due variabili della riga di collegamento e della colonna. Nello specifico, corrispondono al valore del p value. Cliccando sui tre menu a tendina al lato sinistro della heatmap, l'utente è in grado di impostare diverse configurazioni dell'esperimento avendo a disposizione la scelta delle soglie del valore alpha, della dimensione del sample e della percentuale di split tra il train set e il test set.

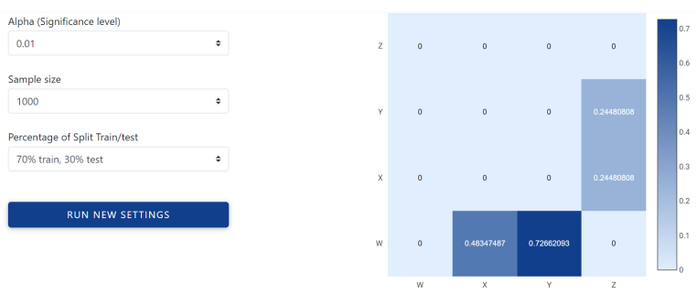

Figure 6: (a) Dropdown menu per esplorare le diverse configurazioni dell'esperimento (b) Mappa di calore rappresentante le risultanti relazioni di causalità

4 Conclusione e sviluppi futuri

Grazie all'utilizzo di questa interfaccia, l'utente è in grado di acquisire una profonda conoscenza dell'approccio proposto da Hoyer ed al.[3]. In particolare l'applicazione offre uno strumento interattivo per l'esplorazione della procedura step-by-step e capace di porre in rilievo gli aspetti essenziali che conducono all'identificazione delle relazioni causali non lineari. Inoltre, mostra gli esperimenti effettuati sulle varie tipologie di dataset (reale e sintetico), accumunate dalla conoscenza della ground truth della direzione della relazione in modo tale da poter confrontare la vera struttura casuale con quella stimata.
Negli sviluppi futuri, sarebbe utile dare la possibilità all'utente di caricare e manipolare i propri dati. Infine, sarebbe interessante poter confrontare visivamente due o più modelli causali in un'unica visualizzazione.

References

[1] Tuan Nhon Dang, Paul Murray, Jillian Aurisano, and Angus Graeme Forbes. Reactionflow: an interactive visualization tool for causality analysis in biological pathways. In *BMC proceedings*, volume 9, page S6. Springer, 2015.

[2] Niklas Elmqvist and Philippas Tsigas. Animated visualization of causal relations through growing 2d geometry. *Information Visualization*, 3(3):154–172, 2004.

[3] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[4] Judea Pearl. *Causality*. Cambridge university press, 2009.

[5] Guillermo Vigueras and Juan A Botia. Tracking causality by visualization of multi-agent interactions using causality graphs. In *International Workshop on Programming Multi-Agent Systems*, pages 190–204. Springer, 2007.

[6] Jun Wang and Klaus Mueller. Visual causality analysis made practical. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 151–161. IEEE, 2017.