Analisi e Mitigazione del Pregiudizio Algoritmico nei Sistemi di Raccomandazione di Contenuti Didattici

Gianni Fenu e Mirko Marras
Dipartimento di Matematica e Informatica, Università di Cagliari, V. Ospedale 72, 09124 Cagliari, Italia
{fenu, mirko.marras}@unica.it

Abstract

I sistemi di raccomandazione sono spesso valutati su quanto accuratamente predicono le preferenze degli utenti. Tuttavia, la letteratura ha mostrato che diversi algoritmi raggiungono una buona accuratezza ma sono caratterizzati da pregiudizi, come suggerire solo elementi popolari. Con attenzione ai sistemi di raccomandazione di contenuti didattici, stiamo studiando come gli algoritmi esistenti propagano i pregiudizi dei dati usati per addestrarli e stiamo sviluppando metodi per mitigarne l'effetto.

1 Contesto di riferimento e motivazioni

I sistemi di raccomandazione mirano a suggerire elementi che potrebbero interessare ad un utente, come quali articoli comprare o quale brani musicali ascoltare. L'apprendimento automatico e la grande mole di dati disponibile rappresentano elementi fondamentali che consentono a tali sistemi intelligenti di apprendere dagli utenti e adattare il loro output alle esigenze e alle preferenze degli utenti. Tuttavia, la letteratura ha mostrato come queste tecnologie, unitamente ai pregiudizi presenti nei dati usati per addestrarle, stiano sollevando nuove sfide etiche e sociali [Hajian *et al.*, 2016]. È apparso evidente che concentrarsi solo sulle preferenze dell'utente ha oscurato altri risultati importanti e i reali benefici che tali sistemi dovrebbero fornire. Correttezza, trasparenza, equità, apertura alla diversità e altre proprietà sul benessere sociale non vengono catturate dalle metriche su cui tipicamente questi modelli personalizzati sono ottimizzati, quali l'accuratezza nel predire le preferenze degli utenti [Bellogín *et al.*, 2017]. Pertanto, valutare come i sistemi di raccomandazione gestiscono i pregiudizi che minano le proprietà citate diventa cruciale.

La formazione online e i sistemi di raccomandazione qui impiegati rappresentano un'interessante area da investigare. Gran parte delle piattaforme e-learning, specialmente quelle che offrono corsi su larga scala, ha attratto molti partecipanti e le loro interazioni hanno generato grandi quantità di dati. La disponibilità di questi dati ha portato ad una crescente diffusione di metodi per l'analisi dell'apprendimento e di nuove opportunità per supportare la didattica online [Drachsler *et al.*, 2015]. Si prevede che il mercato in questo settore cresca da 2,6 miliardi di USD nel 2018 a 7,1 miliardi di USD nel 2023 [MarketsandMarkets, 2018]. Gli emergenti strumenti di personalizzazione e supporto sono considerati una cura contro parte delle attuali problematiche didattiche [Klašnja-Milićević *et al.*, 2017]. Tali soluzioni mirano comunemente a suggerire materiale digitale (slide o video). Recentemente, è stata introdotta la raccomandazione di corsi online. Considerato l'importante ruolo di queste tecnologie nella formazione, il rischio che ogni pregiudizio ignorato possa avere un impatto su una moltitudine di persone aumenta. La rimozione di ogni pregiudizio è ora impraticabile, ma l'individuazione e la mitigazione dovrebbero essere un obiettivo centrale.

2 Attività di ricerca

Il contesto sopra trattato vede il nostro gruppo di ricerca coinvolto nello studio di algoritmi e nello sviluppo di applicazioni con cui fronteggiare i problemi esistenti. Le attività sono articolate sui seguenti assi fortemente interdipendenti di:

- investigazione sul comportamento e sulle prestazioni dei sistemi di raccomandazione, con attenzione alle modalità con cui questi gestiscono e propagano i pregiudizi presenti nei dati impiegati per il loro addestramento;
- formulazione di teorie e algoritmi di raccomandazione capaci di gestire i pregiudizi nei dati, controllando il compromesso con la soddisfazione degli utenti e le finalità didattiche, fondati su studi svolti in contesti reali;
- addestramento, valutazione e validazione degli algoritmi di raccomandazione su grandi ed eterogenee collezioni di dati, raccolte in ambienti di apprendimento digitale, all'interno di contesti applicativi locali e su larga scala;
- sviluppo di prototipi con cui gli esiti della ricerca si traducono in prodotti funzionanti, aventi impatto positivo sul benessere sociale degli attori dell'ecosistema.

Le attività svolte hanno un orientamento interdisciplinare, condividendo competenze teoriche, tecniche e didattiche, al fine di rispecchiare le esigenze attuali del mondo reale.

3 Caso di studio esemplificativo

Questa sezione mostra parte dell'analisi condotta sul comportamento dei sistemi di raccomandazione di corsi online [Boratto *et al.*, 2019]. I sistemi selezionati sono stati addestrati con le valutazioni numeriche (rating) lasciate dagli studenti a fine corso. In aggiunta ad alcune baseline (MostPop,

Random, ItemAvg, UserAvg), sono stati considerati algoritmi applicati in contesti e-learning: alcuni ottimizzano la predizione del valore dei rating (ItemKNN, UserKNN, SVD++, WRMF), altri l'utilità delle liste suggerite (AoBPR, BPR, Hybrid, LDA) [Drachsler *et al.*, 2015]. I dati provengono da un popolare marketplace di corsi online [Dessì *et al.*, 2018]. Per mantenere l'analisi computazionalmente trattabile, sono stati scelti gli studenti che hanno assegnato almeno 10 rating. Il dataset includeva 37K studenti, 600K rating e 30K corsi.

Le liste di corsi suggerite dagli algoritmi sono state comparate in base ai pregiudizi rilevati nei dati usati per l'addestramento. In particolare, si è investigato come il pregiudizio sulla popolarità del corso influenza le liste suggerite e il livello di esposizione del catalogo dei corsi, e come il pregiudizio sulla popolarità delle categorie dei corsi valutate dagli studenti si propaga nelle liste suggerite. Ciascun pregiudizio potrebbe avere implicazioni nell'apprendimento e nell'insegnamento. Di seguito, si evidenzia la parte che mostra l'effetto della popolarità dei singoli corsi sulle liste di corsi suggerite.

Investigare questo effetto appare cruciale. Spesso, si presume che suggerire ciò che è popolare promuova contenuti di qualità ma, non sempre, la popolarità riflette la qualità. Primo, potrebbe esserci un'influenza tra gli utenti e una mancanza di indipendenza. Secondo, le metriche di popolarità potrebbero essere manipolate con recensioni false, per esempio. Terzo, il costo cognitivo di apprendere come valutare la qualità potrebbe portare a corsi con alta popolarità, indipendentemente dalla qualità. Quarto, i corsi meno popolari potrebbero aiutare a capire meglio le preferenze. Quinto, il pregiudizio potrebbe impedire ai nuovi contenuti di emergere e il mercato potrebbe essere dominato da pochi enti o docenti.

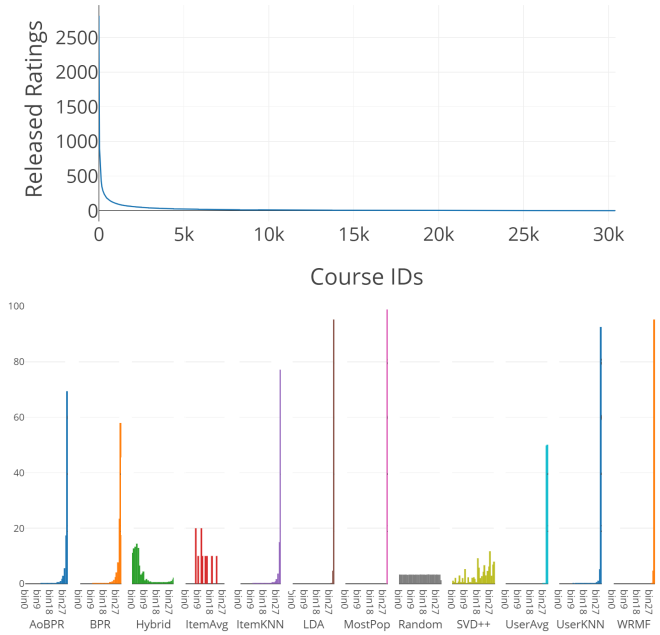

Figura 1: (*In alto*) La distribuzione del numero di rating per corso nel dataset; i corsi sono ordinati per popolarità decrescente. (*In basso*) La distribuzione dei corsi suggeriti in base alla loro popolarità; l'asse X riporta riporta 31 gruppi da 1000 corsi ordinati per crescente popolarità, l'asse Y la rispettiva percentuale di corsi suggeriti.

Figura 1 mostra in alto il pregiudizio sulla popolarità presente nel dataset. La popolarità di un corso nel dataset è calcolata come il numero di rating che esso ha ricevuto. In basso, Figura 1 evidenzia come questo pregiudizio viene propagato nelle liste suggerite. I corsi sono stati ordinati in base al numero di rating nel dataset e suddivisi in gruppi di 1000 corsi; il gruppo più a sinistra contiene i 1000 corsi meno popolari, mentre quelli successivi considerano corsi di crescente popolarità. Per ogni gruppo, viene riportata la percentuale dei corsi ad esso appartenenti e suggeriti dal corrispondente algoritmo. Si noti come quasi tutti gli algoritmi tendono a suggerire corsi del gruppo dei più popolari (quello più a destra). Sebbene Random non soffra del pregiudizio, le sue raccomandazioni non sono ottimizzate secondo alcuna metrica. In BPR, la popolarità del corso sembra essere correlata alla possibilità di essere raccomandato. SVD++ e Hybrid consigliano anche corsi meno popolari. È interessante notare che Hybrid tende a suggerire più corsi poco popolari rispetto a quelli popolari. Complessivamente, nonostante le differenze rispetto all'accuratezza possano essere minime, gli algoritmi hanno mostrato comportamenti diversi e alcune volte indesiderati.

L'analisi completa porta alla luce la necessità di creare sistemi di raccomandazione consapevoli dei pregiudizi nei dati. Nelle attività future, si prevede di investigare altri pregiudizi legati a specifiche categorie di utenti e ai loro attributi (es. genere) e di progettare contromisure contro i pregiudizi rilevati, gestendo il compromesso con la soddisfazione degli utenti.

Ringraziamenti

Le attività sono parzialmente supportate dal Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) nell'ambito del progetto "iLearnTV Anywhere Anytime" (DD n.1937 05.06.2014, CUP F74G14000200008 F19G14000910008).

Riferimenti bibliografici

[Bellogín *et al.*, 2017] A. Bellogín, P. Castells, e I. Cantador. Statistical biases in information retrieval metrics for recommender systems. *Information Retrieval Journal*, 20(6):606–634, 2017.

[Boratto *et al.*, 2019] L. Boratto, G. Fenu, e M. Marras. The effect of algorithmic bias on recommender systems for massive open online courses. In *European Conference on Information Retrieval*. Springer, 2019.

[Dessì *et al.*, 2018] D. Dessì, G. Fenu, M. Marras, e D. Reforgiato. Coco: Semantic-enriched collection of online courses at scale with experimental use cases. In *World Conference on Information Systems and Technologies*, pages 1386–1396. Springer, 2018.

[Drachsler *et al.*, 2015] H. Drachsler, K. Verbert, O. Santos, e N. Manouselis. Panorama of recommender systems to support learning. In *Recommender systems handbook*. Springer, 2015.

[Hajian *et al.*, 2016] S. Hajian, F. Bonchi, e C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *ACM SIGKDD Intern. Conference on Knowledge Discovery and Data Mining*, pages 2125–2126. ACM, 2016.

[Klašnja-Milićević *et al.*, 2017] A. Klašnja-Milićević, B. Vesin, M. Ivanović, Z. Budimac, e L. Jain. Recommender systems in e-learning environments. In *E-Learning Systems*. Springer, 2017.

[MarketsandMarkets, 2018] MarketsandMarkets. *Education and Learning Analytics Market*, 2018. http://bit.ly/2X9e3pO.