Equità Algoritmica dei Modelli di Apprendimento Automatico per la Profilazione Utente e la Personalizzazione

Ludovico Boratto, Gianni Fenu, Mirko Marras
Lab CINI UniCA-DMI, Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari
ludovico.boratto@acm.org, fenu@unica.it, mirko.marras@acm.org

Abstract

I sistemi intelligenti che integrano modelli di profilazione utente e personalizzazione (p.e. i sistemi di raccomandazione e i sistemi biometrici) stanno ricevendo crescente attenzione negli ultimi anni. Sebbene abbiano ormai raggiunto un'accuratezza predittiva elevata, tali modelli non sono immuni da rischi osservati in altri ambiti riguardo la possibile discriminazione di (gruppi di) individui caratterizzati da un certo attributo sensibile. Questo articolo fornisce una panoramica dell'attività di ricerca condotta dal Lab CINI UniCA-DMI sulla progettazione e lo sviluppo della prossima generazione di modelli di profilazione utente e personalizzazione responsabili, con particolare attenzione a come caratterizzarne e garantirne l'equità.

1 Introduzione

I progressi impressionanti nei processi decisionali basati sui dati stanno trasformando la nostra società sempre di più. I sistemi dotati di *intelligenza artificiale* (IA) stanno avendo un impatto su ogni aspetto della nostra vita quotidiana, sia online che offline. Tra i sistemi che stanno ricevendo attenzione negli ultimi anni, quelli che integrano modelli di *profilazione utente* e *personalizzazione* (p.e. i sistemi di raccomandazione e i sistemi biometrici) giocano un ruolo primario nel rendere la fruizione delle piattaforme sovrastanti aderente ai *bisogni* e alla *caratteristiche* del singolo utente. Questa intelligenza automatizzata viene utilizzata in una miriade di piattaforme in diversi domini, dal commercio all'istruzione, dalla sanità ai social media, dall'intrattenimento al reclutamento professionale.

Le considerazioni etiche hanno ricevuto un'attenzione crescente in ambito industriale e accademico, dato che i sistemi basati sull'IA influenzano sempre più ogni aspetto della nostra vita. Quelli sopra menzionati non sono immuni dai rischi osservati in altri ambiti e, negli ultimi anni, l'attenzione sulle insidie nel loro impiego è aumentata. Eppure, nonostante questa attenzione diffusa, gli aspetti legati all'*equità algoritmica* sono ancora poco esplorati nei modelli di profilazione utente e personalizzazione, e il lavoro recente ha sottolineato che i requisiti di progettazione e i metodi per affrontare le sfide relative a questa prospettiva sono dipendenti dal *dominio* e dal *sistema*. Così, per completare le indagini generali, comuni nella letteratura su IA, diventa fondamentale che la ricerca sui modelli di profilazione utente e personalizzazione esplori cosa significhi e come garantire l'equità in questo contesto, per far agire tali modelli in modo *responsabile*.

In risposta a queste necessità, il `Lab CINI UniCA-DMI` sta conducendo ricerche originali di alta qualità e di grande impatto sulla teoria e la pratica della progettazione della prossima generazione di modelli di profilazione utente e personalizzazione responsabili. Il gruppo di ricerca opera su tutte le prospettive relative all'equità, compresi nuovi approcci di *modellazione*, *contromisure*, *linee guida*, *protocolli di valutazione*, e *casi d'uso* nel mondo reale. Ogni contributo è contestualizzato rispetto allo scenario del mondo reale considerato, mostrando come esso supporta le parti interessate. Quest'area di ricerca è ritenuta strategica per migliorare la nostra comprensione a livello di comunità scientifica su un importante tema come quello dell'equità algoritmica di tali modelli, basandoci sulla conoscenza acquisita negli anni.

2 Temi di Ricerca

L'implementazione responsabile dei modelli di profilazione utente e personalizzazione nel mondo reale richiede metodi e processi che mettano le *persone al primo posto*, controllino l'*impatto sociale ed etico* e aumentino la *fiducia degli utenti* nelle tecnologie risultanti. Per promuovere questa visione, il `Lab CINI UniCA-DMI` ha proposto e sta proponendo contributi scientifici che coprono un'estesa lista di prospettive riguardanti l'equità algoritmica.

Raccolta e preparazione di set di dati. Ci occupiamo di studiare l'interazione tra equità e dati sbilanciati (o classi rare), progettare metodi per affrontare sbilanciamenti e disuguaglianze nei dati, creare processi di raccolta che portano a set di dati meno sbilanciati, raccogliere insiemi di dati utili per l'analisi di situazioni non eque, e progettare protocolli di raccolta per insiemi di dati su misura per la ricerca sull'equità.

Progettazione e sviluppo di contromisure. Conduciamo diverse attività legate a formalizzare e operazionalizzare il concetto di equità, condurre analisi esplorative che scoprono nuovi tipi di pregiudizi algoritmici legati all'equità, progettare trattamenti che mitigano tali pregiudizi in fase di pre-/in-/post-elaborazione, ideare metodi per spiegare i pregiudizi

identificati, studiare il ragionamento causale e controfattuale legati ai pregiudizi algoritmici sull'equità.

Definizione di protocolli di valutazione. Eseguiamo studi di auditing rispetto all'equità algoritmica, conduciamo studi sperimentali quantitativi su situazioni non eque, definiamo metriche oggettive che considerano l'equità e i relativi pregiudizi algoritmici, formuliamo protocolli specifici per valutare i modelli esistenti, valutiamo le strategie di mitigazione in domini inesplorati, analizziamo efficienza, riproducibilità e scalabilità dei metodi di mitigazione.

Esplorazione di casi di studio. Le attività vengono svolte su sistemi applicativi specifici (p.e. sistemi di raccomandazione e sistemi biometrici) e su domini specifici (p.e. piattaforme di e-commerce, ambienti educativi, siti web di intrattenimento, piattaforme di notizie, biblioteche digitali e portali di lavoro).

3 Progetti, Applicazioni, ed Iniziative

Seguendo i temi di ricerca sopra descritti, il `Lab CINI UniCA-DMI` sta portando avanti molteplici progetti, applicazioni, ed iniziative sull'equità algoritmica dei modelli di profilazione utente e personalizzazione. In quanto segue, descriveremo gli aspetti salienti di alcune di queste attività.

Fair voice biometrics. I sistemi di riconoscimento vocale stanno giocando un ruolo chiave nelle moderne applicazioni online (p.e. gli assistenti vocali intelligenti). Questo progetto ha l'obiettivo di analizzare la suscettibilità di questi sistemi alla discriminazione, considerando gruppi demografici caratterizzati da attributi sensibili. Stiamo studiando come le metriche di equità di gruppo esistenti si relazionano con le impostazioni di bilanciamento del set di dati impiegato per addestrare i modelli di riconoscimento del parlatore. Conduciamo questa analisi operazionalizzando diverse definizioni di equità e monitorandole sotto varie impostazioni di bilanciamento dei dati. Gli esperimenti eseguiti su architetture neurali profonde allo stato dell'arte, valutate su un set di dati che include gruppi basati su sesso/età, mostrano che il bilanciamento della rappresentazione dei gruppi ha un impatto positivo sull'equità tra i gruppi demografici e che l'attrito tra sicurezza, usabilità e accuratezza dipende dalla metrica di accuratezza e dalla soglia di riconoscimento [Fenu *et al.*, 2020a; Fenu *et al.*, 2020b; Fenu *et al.*, 2021]. Set di dati, codice sorgente e modelli pre-addestrati sono disponibili online[1].

Fair personalized student support. Le piattaforme educative online stanno giocando un ruolo primario nel mediare il successo delle carriere degli individui. Nei servizi di personalizzazione dell'esperienza di apprendimento degli studenti, diventa essenziale garantire che essi ricevano opportunità di apprendimento eque in termini di qualità, secondo i valori della piattaforma ed il contesto formativo. L'importanza di garantire l'uguaglianza in qualità delle opportunità di apprendimento è stata ben studiata in passato, tuttavia come essa possa essere attuata negli ecosistemi di apprendimento attraverso i sistemi di raccomandazione è ancora poco esplorato. In questo progetto, formalizziamo i principi educativi che modellano le proprietà di apprendimento delle raccomandazioni e metriche di equità che li combinano, per monitorare l'uguaglianza in qualità delle opportunità raccomandate. Esploriamo poi le opportunità suggerite dai sistemi di raccomandazione agli studenti, scoprendo disuguaglianze sistematiche. Per ridurne l'effetto, proponiamo approcci che bilanciano la personalizzazione e l'uguaglianza in qualità delle opportunità raccomandate [Boratto *et al.*, 2019; Marras *et al.*, 2021].

Geographic provider fairness in recommendation. In un sistema di raccomandazione che suggerisce elementi agli utenti, si dà una certa esposizione ai fornitori dietro gli elementi raccomandati. Il sistema offre infatti una possibilità agli elementi di quei fornitori di essere raggiunti e consumati dagli utenti. Quindi, a seconda di come le liste di raccomandazione sono modellate, l'esperienza dei fornitori sottoraccomandati può essere influenzata. Per studiare questo fenomeno, ci concentriamo sulla raccomandazione di film e libri, as esempio, e arricchiamo i set di dati con il continente di produzione di ciascun elemento. Usiamo questi dati per caratterizzare gli squilibri nella distribuzione della provenienza degli elementi prodotti (squilibrio geografico). Per valutare se i sistemi di raccomandazione generano un impatto disparato e (s)favoriscono un gruppo, dividiamo gli elementi in gruppi, in base al loro continente di produzione, e caratterizziamo quanto ogni gruppo è rappresentato nei dati. Poi, eseguiamo sistemi di raccomandazione all'avanguardia e misuriamo la visibilità e l'esposizione data a ciascun gruppo. Osserviamo disparità che favoriscono i gruppi più rappresentati. Introduciamo l'equità con approcci che regolano la quota di raccomandazioni date agli elementi prodotti in un continente (visibilità) e le posizioni in cui essi sono classificati nella lista di raccomandazione (esposizione) [Gómez *et al.*, 2021a; Gómez *et al.*, 2021b; Gómez *et al.*, 2022a; Gómez *et al.*, 2022b].

Equal treatment of items in personalized rankings. I sistemi di raccomandazione sono addestrati su dati che veicolano le preferenze degli utenti. Tuttavia, queste sono spesso distribuite in modo non uniforme tra gli elementi da suggerire. Di conseguenza, questi sistemi possono finire per suggerire progressivamente gli elementi popolari più di quelli di nicchia, anche quando questi ultimi sarebbero di interesse per gli utenti. Questo può ostacolare diverse qualità fondamentali delle liste raccomandate (p.e. novità, copertura, diversità), con un impatto sul futuro successo della piattaforma stessa. In questo progetto, formalizziamo nuove metriche che quantificano quanto un sistema di raccomandazione tratti equamente gli elementi lungo la coda della popolarità. Caratterizziamo le raccomandazioni di algoritmi rappresentativi per mezzo delle metriche proposte, e mostriamo che la probabilità di essere raccomandati è distorta dalla popolarità dell'elemento. Per promuovere un trattamento più equo, proponiamo approcci volti a minimizzare la correlazione distorta tra la rilevanza dell'elemento utente e la popolarità dell'elemento [Boratto *et al.*, 2021c; Boratto *et al.*, 2021b; Boratto *et al.*, 2021d].

Ulteriori iniziative di divulgazione scientifica. I membri del `Lab CINI UniCA-DMI` ricoprono un ruolo di primo piano

[1] https://mirkomarras.github.io/fair-voice/

nel presiedere workshop tenuti in combinazione con conferenze internazionali di alto livello, tra cui ECIR BIAS (2020 [Boratto *et al.*, 2020], 2021 [Boratto *et al.*, 2021a], 2022[2]), WSDM L2D (2021[3]), ICCV RPRMI (2021[4]). Tengono tutorial scientifici su tematiche riguardanti l'equità algoritmica, tra cui quelli erogati a UMAP (2020 [Boratto e Marras, 2020]), ICDM (2020[5]), WSDM (2021[6]), ICDE (2021 [Boratto e Marras, 2021]), e ECIR (2021[7]). Infine, sono anche editori ospiti per numeri speciali su riviste del settore come Information Processing & Management [Boratto *et al.*, 2022] e Future Generation Computer Systems [Bonnin *et al.*, 2022].

4 Sfide e Prospettive

Nonostante i recenti progressi in questo campo, caratterizzare e garantire l'*equità algoritmica* nei modelli di *profilazione utente* e *personalizzazione* presenta ancora molteplici sfide.

Molte di queste sono legati alle *specificità del dominio*. Per esempio, gli attori coinvolti (p.e. consumatori e produttori nelle piattaforma di commercio elettronico) hanno diversi e contrastanti bisogni. Fare in modo che i modelli sviluppati tengano conto dei *bisogni e degli interessi di tutti gli attori* in maniera responsabile richiede ulteriore ricerca, magari attraverso approcci multidisciplinari (p.e. per collegare giustizia ed equità). Sintetizzare una *definizione di pregiudizio o di equità algoritmica* è impegnativo così come lo è creare un vocabolario comune per riconoscere diversi tipi di pregiudizi e/o inequità. Ulteriori sfide solo legate alla *mancanza di dati* per caratterizzare i fenomeni di equità algoritmica con sufficiente profondità (specialmente set di dati che riportano gli attributi sensibili degli utenti), e ci sono forme di pregiudizio algoritmico che non sono state studiate nella letteratura.

Da un punto di vista operativo, *misurare* e *operazionalizzare* concetti legati al pregiudizio o all'equità algoritmica include diverse sfide. Per esempio, non è ancora chiaro come poter *mitigare più forme di inequità* allo stesso tempo, e come piccoli cambiamenti nel processo di sviluppo possano fare un'enorme *differenza sull'impatto* dei modelli sugli utenti. Focalizzare maggiormente la ricerca e lo sviluppo sull'*applicazione del mondo reale*, includendo gli *utenti nel processo di valutazione* dei modelli rappresenta un'altra prospettiva da esplorare. Inoltre, quando si mitigano dei pregiudizi algoritmici, di solito, si osservano delle perdite su altre proprietà (p.e. accuratezza). Come poter mitigare l'inequità *senza compromettere le altre proprietà* sarà oggetto di ricerca futura da parte del nostro laboratorio.

Infine, diverse sono le sfide legate alla *valutazione* dei modelli dal punto di vista della loro equità di trattamento verso le categorie di utenti per le quali sono state sviluppate. Spesso, non gli *attributi sensibili degli utenti non sono inclusi* nei dati raccolti. Non è ancora chiaro come dovremmo *selezionare un approccio rispetto a un altro* (p.e. scegliere tra differenti definizioni di equità). Non sempre, inoltre, le metriche e gli esperimenti condotti offline *non offrono una buona stima dell'impatto reale* dei modelli quando queste verranno applicate in piattaforma esistenti. Considerare sia gli obiettivi di equità che l'*efficienza* è tutt'ora un altro aspetto da esplorare.

Il `Lab CINI UniCA-DMI` si sta occupando di affrontare tali sfide e prospettive al fine di sviluppare la prossima generazione di modelli di profilazione utente e personalizzazione responsabili, con attenzione a come caratterizzarne e garantirne l'equità l'algoritmica.

[2]https://biasinrecsys.github.io/ecir2022/

[3]https://mirkomarras.github.io/l2d-wsdm2021/

[4]https://rprmiworkshop.github.io/iccv2021/

[5]https://biasinrecsys.github.io/icdm2020/

[6]https://biasinrecsys.github.io/wsdm2021/

[7]https://biasinrecsys.github.io/ecir2021-tutorial

Riferimenti bibliografici

[Bonnin *et al.*, 2022] Geoffray Bonnin, Danilo Dessì, Gianni Fenu, Martin Hlosta, Mirko Marras, e Harald Sack. Guest editorial of the FGCS special issue on advances in intelligent systems for online education. *Future Gener. Comput. Syst.*, 127:331–333, 2022.

[Boratto *et al.*, 2019] Ludovico Boratto, Gianni Fenu, e Mirko Marras. The effect of algorithmic bias on recommender systems for massive open online courses. In *Proc. 41st European Conference on IR Research - ECIR 2019*, volume 11437 of *Lecture Notes in Computer Science*, pages 457–472. Springer, 2019.

[Boratto *et al.*, 2020] Ludovico Boratto, Stefano Faralli, Mirko Marras, e Giovanni Stilo, editors. *Bias and Social Aspects in Search and Recommendation - First International Workshop on Algorithmic Bias in Search and Recommendation - BIAS 2020*, volume 1245 of *Communications in Computer and Information Science*. Springer, 2020.

[Boratto *et al.*, 2021a] Ludovico Boratto, Stefano Faralli, Mirko Marras, e Giovanni Stilo, editors. *Advances in Bias and Fairness in Information Retrieval - Second International Workshop on Algorithmic Bias in Search and Recommendation - BIAS 2021*, volume 1418 of *Communications in Computer and Information Science*. Springer, 2021.

[Boratto *et al.*, 2021b] Ludovico Boratto, Gianni Fenu, e Mirko Marras. Combining mitigation treatments against biases in personalized rankings: Use case on item popularity. In *Proc. 11th Italian Information Retrieval Workshop - IIR 2021*, volume 2947 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[Boratto *et al.*, 2021c] Ludovico Boratto, Gianni Fenu, e Mirko Marras. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Inf. Process. Manag.*, 58(1):102387, 2021.

[Boratto *et al.*, 2021d] Ludovico Boratto, Gianni Fenu, e Mirko Marras. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Model. User Adapt. Interact.*, 31(3):421–455, 2021.

[Boratto *et al.*, 2022] Ludovico Boratto, Stefano Faralli, Mirko Marras, e Giovanni Stilo. Guest editorial of the IPM special issue on algorithmic bias and fairness in search and recommendation. *Inf. Process. Manag.*, 59(1):102791, 2022.

[Boratto e Marras, 2020] Ludovico Boratto e Mirko Marras. Hands on data and algorithmic bias in recommender systems. In *Pro. 28th ACM Conference on User Mode-*

ling, Adaptation and Personalization - UMAP 2020, pages 388–389. ACM, 2020.

[Boratto e Marras, 2021] Ludovico Boratto e Mirko Marras. Countering bias in personalized rankings : From data engineering to algorithm development. In *Proc. 37th IEEE International Conference on Data Engineering - ICDE 2021*, pages 2362–2364. IEEE, 2021.

[Fenu *et al.*, 2020a] Gianni Fenu, Hicham Lafhouli, e Mirko Marras. Exploring algorithmic fairness in deep speaker verification. In *Proc. 20th International Conference on Computational Science and Its Applications - ICCSA 2020*, volume 12252 of *Lecture Notes in Computer Science*, pages 77–93. Springer, 2020.

[Fenu *et al.*, 2020b] Gianni Fenu, Giacomo Medda, Mirko Marras, e Giacomo Meloni. Improving fairness in speaker recognition. In *Proc. European Symposium on Software Engineering - ESSE 2020*, pages 129–136. ACM, 2020.

[Fenu *et al.*, 2021] Gianni Fenu, Mirko Marras, Giacomo Medda, e Giacomo Meloni. Fair Voice Biometrics: Impact of Demographic Imbalance on Group Fairness in Speaker Recognition. In *Proc. Interspeech 2021*, pages 1892–1896, 2021.

[Gómez *et al.*, 2021a] Elizabeth Gómez, Ludovico Boratto, e Maria Salamó. Disparate impact in item recommendation: A case of geographic imbalance. In *Proc. 43rd European Conference on IR Research - ECIR 2021*, volume 12656 of *Lecture Notes in Computer Science*, pages 190–206. Springer, 2021.

[Gómez *et al.*, 2021b] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, e Mirko Marras. The winner takes it all: Geographic imbalance and provider (un)fairness in educational recommender systems. In *Proc. 44th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 2021*, pages 1808–1812. ACM, 2021.

[Gómez *et al.*, 2022a] Elizabeth Gómez, Ludovico Boratto, e Maria Salamó. Provider fairness across continents in collaborative recommender systems. *Inf. Process. Manag.*, 59(1):102719, 2022.

[Gómez *et al.*, 2022b] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, e Guilherme Ramos. Enabling cross-continent provider fairness in educational recommender systems. *Future Gener. Comput. Syst.*, 127:435–447, 2022.

[Marras *et al.*, 2021] Mirko Marras, Ludovico Boratto, Guilherme Ramos, e Gianni Fenu. Equality of learning opportunity via individual fairness in personalized recommendations. *International Journal of Artificial Intelligence in Education*, pages 1–49, 2021.