Integrazione automatica di dati biomedicali eterogenei

Marco Aiello[1], **Andrea Apicella**[2], **Piero A. Bonatti**[2], **Anna Corazza**[2], **Francesco Isgrò**[2], **Iliana M. Petrova**[2], **Roberto Prevete**[2], **Daniel Riccio**[2], **Luigi Sauro**[2], **Guglielmo Tamburrini**[2]

[1] IRCCS SDN, Napoli

[2] Laboratorio di Artificial Intelligence, Privacy and Applications (AIPA)
DIETI, Università degli Studi di Napoli Federico II
http://aipa.dieti.unina.it

Abstract

Nell'ambito biomedicale, l'applicazione di tecniche di intelligenza artificiale a dati eterogenei mediante l'integrazione di conoscenza ontologica probabilistica presenta interessanti sfide soprattutto per quel che riguarda la necessità di costruire spiegazioni di quanto ricavato e di rispettare la privatezza dei dati e principi etici condivisi.

1 Introduzione

I sistemi informativi di supporto alla diagnosi medica sono per loro natura eterogenei. In tale caratteristica risiedono, al contempo, l'enorme potenziale ancora inespresso e la commisurata difficoltà nell'armonizzare fonti di informazione fra loro così diverse. Basti pensare al fatto che la storia clinica del paziente include una parte testuale, rappresentata da cartelle cliniche e referti medici, alla quale si affiancano dati numerici risultanti dalle analisi di laboratorio e dati multimediali (immagini e video) prodotti dalla diagnostica per immagini e dall'istopatologia al microscopio, solo per citarne alcune. La diagnosi prodotta dal medico nasce, quindi, dall'integrazione di tutte le informazioni disponibili alla luce di competenze acquisite tramite lo studio di trattati scientifici e dall'esperienza professionale sul campo. I notevoli progressi fatti nei singoli settori inerenti il trattamento dei dati, l'estrazione di caratteristiche salienti e la conseguente organizzazione e gestione della conoscenza, rappresentano un punto di partenza ottimale per affrontare la principale sfida attuale: l'integrazione multi-modale di fonti eterogenee di informazione a supporto della diagnosi medica. Tale processo passa per aspetti fondamentali legati al paziente, quali il rispetto della privacy e dei principi etici, o ai sistemi di supporto realizzati, come l'"explainability" dei processi decisionali intrinseci al sistema.

Alla luce delle pregresse esperienze maturate nella partecipazione in progetti come il PON SmartHealth diretto da Engineering Ingegneria Informatica S.p.A. e NoemaLife S.p.A., il gruppo di ricerca si pone come obiettivo la progettazione di sistemi innovativi di supporto alla diagnostica che mirino all'integrazione multimodale dei dati, senza perdere di vista aspetti non secondari quali l'etica, la privacy e l'interpretabilità delle decisioni. Il PON SmartHealth si è posto come target la creazione di un'infrastruttura tecnologica per lo sviluppo di servizi nell'area della salute e del benessere. All'interno del progetto, il gruppo è stato principalmente coinvolto su aspetti legati alla gestione delle cartelle cliniche, sia per l'estrazione di entità e relazioni dalle sezioni in linguaggio naturale [Alicante *et al.*, 2016], sia per il rafforzamento della privacy mediante controllo degli accessi semantico a prova di inferenza [Bonatti and Sauro, 2013; Bonatti *et al.*, 2015]. Le questioni relative al trattamento dei dati sensibili e alla conformità con la GDPR vengono attualmente approfondite nel progetto H2020 SPECIAL (www.specialprivacy.eu), mediante la gestione del consenso dei data subject vista come problema di knowledge management, e attraverso forme *privacy-preserving* di data mining. A queste esperienze, si aggiungono ulteriori competenze maturate nell'ambito del medical imaging, attraverso la collaborazione con importanti realtà ospedaliere presenti sul territorio, fra le quali si annoverano: i) analisi di immagini ecografiche del feto per la per la segmentazione e misurazione della plica nucale [Catanzariti *et al.*, 2009; Anzalone *et al.*, 2013], ii) segmentazione di vasi sanguigni in immagini della retina [Frucci *et al.*, 2016], iii) segmentazione del fondo oculare per la diagnosi della retinite pigmentosa [Brancati *et al.*, 2018a], iv) analisi di immagini istopatologiche per la diagnosi del cancro al seno [Brancati *et al.*, 2018b], v) analisi di sequenze video del letto ungueale [Isgrò *et al.*, 2013] e vi) l'analisi di immagini a immunofluorescenza da microscopio [Riccio *et al.*, 2019]; vi) integrazione di informazione derivata da imaging diagnostico e biologia molecolare (radiogenomica) [Incoronato *et al.*, 2017]; vii) calcolo di matrici di connettività cerebrale da dati di imaging multimodale [Aiello *et al.*, 2016].

2 Attività di ricerca

Nel contesto delle applicazioni per il supporto alla diagnostica, sono ancora molti i problemi aperti. Primo fra tutti, l'integrazione di sorgenti di informazione eterogenee. Le tecniche classiche di estrazione dell'informazione vanno adattate al particolare dominio biomedicale, dove occorre concentrarsi su particolari tipi di entità e relazioni che riguardano tipicamente patologie, farmaci, esami, etc. [Alam *et al.*, 2016]. A titolo di esempio, l'informazione tratta da documenti testuali o da test genetici, può essere utilizzata per migliorare la classificazione automatica di oggetti all'interno di immagini biomedicali o per raffinare il risultato della diagnosi finale.

Nel caso specifico della retinite pigmentosa, è ormai appurato che tale patologia abbia origine dalla mutazione di un gruppo specifico di geni, la quale porta all'alterazione del tessuto della retina con la conseguente formazione di pigmenti sia nella zona centrale, che periferica del fondo oculare. L'analisi congiunta dei risultati dei test genetici e delle immagini del fondo oculare offre il vantaggio di poter classificare con una più elevata precisione la natura e lo stadio della patologia.

In secondo luogo, il gruppo di ricerca ha fatto propria la convinzione che l'integrazione di informazioni in un contesto multi-modale, non può prescindere dalla conoscenza a priori resa disponibile al personale medico attraverso lo studio di trattati medici. Tale conoscenza può essere organizzata in ontologie, che si stanno rendendo disponibili in modo sempre più completo e che possono essere costruite, oltre che manualmente, in modo automatico o semi-automatico a partire da testi specialistici. Tuttavia, risulta spesso limitante la necessità di imporre tale conoscenza in modo netto, laddove una descrizione probabilistica risulta spesso più opportuna [Apicella *et al.*, 2017].

Un ulteriore aspetto, che diventa assolutamente irrinunciabile in ambiti delicati come quello biomedicale è rappresentato dalla interpretabilità del processo decisionale alla base di un procedimento automatico di estrazione e sintesi dell'informazione. Ricostruire e spiegare in maniera esauriente le decisioni e i comportamenti degli attuali sistemi [Apicella *et al.*, 2019], basati su tecniche di apprendimento automatico, è molto spesso un compito arduo. In tal senso, il gruppo di ricerca ha focalizzato la propria attenzione sull'uso di ontologie di dominio, possibilmente probabilistiche, benché la conoscenza da esse fornita tipicamente non sia esaustiva e vada integrata con tecniche che permettano di "recuperare" e fornire le strutture esplicative richieste da un essere umano o da un altro sistema autonomo. A riprova della fondatezza di tale convinzione, si osserva che l'ideazione e messa a punto di tali tecniche è il principio fondante della *eXplainable Artificial Intelligence* (XAI).

La generazione di spiegazioni comprensibili per un essere umano, affrontato come problema scientifico e tecnologico nell'ambito della XAI, ha anche una notevole rilevanza di carattere etico-giuridico. Ottenere spiegazioni di classificazioni e decisioni automatiche è importante per la protezione della dignità e dell'autonomia delle persone. Basta ricordare qui l'art. 22 del GDPR, nel quale si afferma il diritto della persona di contestare una decisione automatica che lo riguarda interpellando il gestore del sistema decisionale, il quale, a sua volta, deve porsi il problema della spiegazione dei comportamenti del sistema, e cioè di capire il perché della classificazione/decisione contestata. Inoltre, avere una buona spiegazione in caso di comportamenti indesiderati del sistema può contribuire a migliorare gli interventi correttivi da effettuare anche in ottemperanza al codice etico e di condotta professionale dell'ingegnere informatico.

Riferimenti bibliografici

[Aiello *et al.*, 2016] M. Aiello, C. Cavaliere, and M. Salvatore. Hybrid pet/mr imaging and brain connectivity. *Frontiers in Neuroscience*, 10:64, 2016.

[Alam *et al.*, 2016] F. Alam, A. Corazza, A. Lavelli, and R. Zanoli. A knowledge-poor approach to chemical-disease relation extraction. *Database*, 2016, 2016.

[Alicante *et al.*, 2016] A. Alicante, A. Corazza, F. Isgrò, and S. Silvestri. Unsupervised entity and relation extraction from clinical records in italian. *Comp. in Bio. and Med.*, 72:263–275, 2016.

[Anzalone *et al.*, 2013] A. Anzalone, G. Fusco, F. Isgrò, et al. A system for the automatic measurement of the nuchal translucency thickness from ultrasound video stream of the foetus. In *Proceedings of the 26th CBMS*, pages 239–244, 2013.

[Apicella *et al.*, 2017] A. Apicella, A. Corazza, F. Isgrò, and G. Vettigli. Exploiting context information for image description. In *Proc. ICIAP*, pages 320–331, 2017.

[Apicella *et al.*, 2019] A. Apicella, F. Isgrò, et al. Explaining classification systems using sparse dictionaries. In *Proc ESANN*, 2019.

[Bonatti and Sauro, 2013] P.A. Bonatti and L. Sauro. A confidentiality model for ontologies. In *Proc. of The Semantic Web - ISWC 2013*, pages 17–32, 2013.

[Bonatti *et al.*, 2015] P.A. Bonatti, I.M. Petrova, and L. Sauro. Optimized construction of secure knowledge-base views. In *Proc. DL 2015*, 2015.

[Brancati *et al.*, 2018a] N. Brancati, M. Frucci, D. Gragnaniello, D. Riccio, V. Di Iorio, and L. Di Perna. Automatic segmentation of pigment deposits in retinal fundus images of retinitis pigmentosa. *Computerized Medical Imaging and Graphics*, 66:73–81, 2018.

[Brancati *et al.*, 2018b] N. Brancati, M. Frucci, and D. Riccio. Multi-classification of breast cancer histology images by using a fine-tuning strategy. In *Image Analysis and Recognition*, pages 771–778, 2018.

[Catanzariti *et al.*, 2009] E. Catanzariti, G. Fusco, F. Isgrò, S. Masecchia, R. Prevete, and M. Santoro. A semi-automated method for the measurement of the fetal nuchal translucency in ultrasound images. In *Proceedings ICIAP*, pages 613–622, 2009.

[Frucci *et al.*, 2016] M. Frucci, D. Riccio, G. Sanniti di Baja, and L. Serino. Severe: Segmenting vessels in retina images. *Pattern Recognition Letters*, 82:162 – 169, 2016. An insight on eye biometrics.

[Incoronato *et al.*, 2017] M. Incoronato, M. Aiello, et al. Radiogenomic analysis of oncological data: A technical survey. *International Journal of Molecular Sciences*, 18(4), 4 2017.

[Isgrò *et al.*, 2013] F. Isgrò, F. Pane, et al. Segmentation of nailfold capillaries from microscopy video sequences. In *Proc. of the 26th CBMS*, pages 227–232, June 2013.

[Riccio *et al.*, 2019] D. Riccio, N. Brancati, M. Frucci, and D. Gragnaniello. A new unsupervised approach for segmenting and counting cells in high-throughput microscopy image sets. *IEEE Journal of Biomedical and Health Informatics*, 23(1):437–448, Jan 2019.