Machine learning per Health & Medicine

Gianvito Pio, Michelangelo Ceci, Paolo Mignone, Emanuele Pio Barracchia e Donato Malerba

Laboratorio KDDE – Knowledge and Discovery Data Engineering
Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro
Via Orabona 4, I-70125 Bari, Italy
gianvito.pio@uniba.it, michelangelo.ceci@uniba.it, paolo.mignone@uniba.it,
emanuele.barracchia@uniba.it, donato.malerba@uniba.it

Abstract

In questo contributo si descrivono le attività attualmente in corso ed alcuni recenti risultati ottenuti dal gruppo di ricerca KDDE nello sviluppo di metodi di machine learning nell'ambito Health e Medicine. La ricerca condotta riguarda innanzitutto l'analisi di dati relativi alla genomica, con l'obiettivo principale di investigare relazioni tra diverse entità biologiche (geni, microRNA, ecc.). Inoltre, si sta ampliando lo spettro di intervento considerando possibili relazioni tra entità biologiche e patologie. A tale proposito si stanno studiando le relazioni tra il microbiota umano e le patologie, nell'ambito del progetto UE CA18131 ML4Microbiome. Recenti attività si stanno focalizzando sullo sviluppo di strumenti per la prevenzione, il monitoraggio e la diagnosi, personalizzati sull'individuo. Tali attività di ricerca sono coerenti con gli obiettivi di un progetto di ricerca recentemente finanziato dal piano PON ricerca (area di specializzazione "Tecnologia per gli ambienti della vita").

1 Introduzione

I trend demografici osservabili nella società moderna mostrano un'impennata della curva di invecchiamento che, sebbene possa essere considerato un risultato positivo, costituisce anche uno dei temi di discussione sociali e sanitari che necessitano di essere presi seriamente in considerazione dalle moderne Nazioni. Infatti, l'allungamento della speranza di vita derivante dalle migliori condizioni di vita e dai significativi passi in avanti compiuti dalla medicina, fa stimare che, entro il 2025, oltre il 20% dei cittadini europei sarà over 65.

Il problema principale che si pone, in questo situazione, deriva dalla necessità di definire politiche che garantiscano la sostenibilità della spesa, senza compromettere il diritto alla salute, puntando sempre più a migliorare l'efficacia e l'efficienza del sistema socio-sanitario.

A tal riguardo, l'utilizzo di tecnologie e servizi ICT rappresenta uno mezzo fondamentale su due fronti: *i)* supportare le attività di ricerca in biologia e medicina, tramite metodi di predizione basati su machine learning; *ii)* sviluppare sistemi, strumenti e servizi che rendano le attività giornaliere di prevenzione, monitoraggio e diagnosi, sempre più automatizzate, efficienti e personalizzate sull'individuo.

Riguardo al primo fronte, è rilevante il contributo apportato dalle tecnologie basate su Next Generation Sequencing (NGS) [Reis-Filho, 2009], che hanno permesso la raccolta di grandi moli di dati di natura biologica analizzabili tramite metodi computazionali. Riguardo al secondo fronte, invece, ci si sta muovendo sempre più verso l'introduzione di servizi ICT disegnati sull'esigenza del singolo individuo, sia attraverso l'analisi di stream di dati provenienti da sensori (raccolti, ad esempio, tramite wearable devices), sia tramite l'utilizzo di modelli appresi a partire da dati relativi a casi o situazioni precedentemente osservati in altri individui o nello stesso individuo, in momenti diversi della propria vita.

2 Machine learning per la ricerca in biologia e medicina

Molte delle recenti scoperte in ambito biologico e medico sono state significativamente supportate da tecnologie Next Generation Sequencing (NGS) e da metodi computazionali, che hanno portato ad una migliore comprensione dei meccanismi biologici di diversi organismi e, in particolare, ad una migliore consapevolezza delle possibili relazioni tra entità biologiche e malattie.

A tal riguardo, nel laboratorio KDDE la ricerca si è inizialmente focalizzata sull'individuazione di reti di interazioni geniche e di interazioni tra microRNA e geni. Da un punto di vista metodologico, sono stati sviluppati algoritmi di biclustering [Pio *et al.*, 2013], approcci basati sul meta-learning [Pio *et al.*, 2014], in grado cioè di sfruttare e combinare le predizioni prodotte da diversi metodi di predizione, e approcci basati sul multi-view learning [Ceci *et al.*, 2015]. Infine, è stato sviluppato Comirnet [Pio *et al.*, 2015][1], una applicazione web che consente di interrogare i risultati ottenuti, che rappresenta uno strumento utile per i ricercatori dell'area.

I lavori proposti hanno anche portato alla scoperta di interazioni interessanti, grazie anche alla collaborazione con l'Istituto di Tecnologie Biomediche (ITB - CNR, Bari), tra geni coinvolti nello sviluppo di diverse tipologie di carcinoma nell'uomo (ad esempio, miR-17-92, miR-106b-25 and miR-106a-363 in [Pio *et al.*, 2013]).

[1] http://comirnet.di.uniba.it

La complessità delle reti di interazione, e la frequente individuazione di falsi positivi da parte di metodi di predizione proposti in letteratura, ha spinto il gruppo di ricerca allo sviluppo di un metodo [Pio *et al.*, 2017] in grado di analizzare e cogliere attività di regolazione indirette, al fine di ridurre il numero di falsi positivi (solitamente dovuti a fenomeni di causa comune o effetto comune [Korb e Nicholson, 2010]).

Inoltre, spinti dalla necessità di prendere in considerazione numerose altre entità coinvolte, oltre a quelle oggetto di studio, ci si è focalizzati sullo sviluppo di metodi in grado di analizzare reti di interazioni eterogenee, ossia contenenti entità (biologiche e non) di tipo diverso, in relazione tra loro. In particolare, si è dapprima sviluppato un metodo per il clustering e la classificazione in reti eterogenee [Pio *et al.*, 2018], per poi progettare una versione specifica per la predizione di interazioni tra non-coding RNAs e malattie [Barracchia *et al.*, 2017]. I risultati preliminari mostrano una migliore accuratezza nell'individuazione di nuove interazioni, che possono essere considerate un solido suggerimento per definire il focus di studi in laboratorio.

Infine, di recente il gruppo KDDE ha investigato un'ulteriore strada, che sfrutta tecniche di transfer learning. Queste tecniche sono in genere utilizzate per apprendere modelli in domini in cui la quantità di dati disponibile non è sufficiente, e si cerca di sfruttare la conoscenza relativa ad altri domini affini. In questo specifico caso si è investigata la possibilità di utilizzare la conoscenza relativa a reti di regolazione genica di un organismo per migliorare l'accuratezza di ricostruzione della rete di un altro organismo. In particolare, è stato sperimentata la ricostruzione della rete di regolazione genica umana, sfruttando quella del topo [Mignone e Pio, 2018].

Nell'ambito di questa linea di ricerca, il KDDE è anche coinvolto nel progetto COST "Statistical and machine learning techniques in human microbiome studies (ML4Microbiome)". In particolare, ci si occupa di studiare e valutare lo stato dell'arte, in termini di metodi Machine Learning, nonchè di valutare la robustezza e l'appropriatezza di nuovi metodi proposti in letteratura per la ricerca sul microbiota umano. Inoltre, sulla base di tali risultati, ci si occuperà di definire le aree di intervento prioriarie, per lo sviluppo di nuovi metodi di Machine Learning che possano risolvere specifiche problematiche, e di definire best practices per l'applicazione del Machine Learning per lo studio del microbiota.

3 Machine learning per monitoraggio e prevenzione personalizzati

Nell'ambito dello sviluppo di sistemi e strumenti per prevenzione, monitoraggio e diagnosi, personalizzate sull'individuo, il gruppo KDDE è attualmente coinvolto nelle attività del progetto PON TALIsMAn - (Tecnologie di Assistenza personALizzata per il Miglioramento della quAlità della vitA). In particolare, il KDDE è coinvolto nella definizione di strumenti di analytics finalizzati all'apprendimento di pattern comportamentali utili a riconoscere situazioni anomale, e all'arricchimento del profiling sanitario considerato nel trattamento riabilitativo. In particolare, si studiano approcci di pattern discovery, di supporto alla scoperta di sequenze statisticamente interessanti di attività in grado di rappresentare protocolli di trattamento ottimali, o possibili anomalie o situazioni critiche. Inoltre, il gruppo KDDE lavora alla progettazione di sistemi di monitoraggio di bio-segnali del paziente e di parametri ambientali. In particolare, ci si occupa di realizzare strumenti di analisi che, utilizzando dati del paziente raccolti in tempo reale, siano in grado di rilevare anomalie rispetto a informazioni storiche del paziente, inclusi i modelli di comportamento precedentemente descritti. Tali strumenti forniranno indicazioni decisionali per il paziente anziano e/o per il suo care giver, sulle azioni da porre in atto al verificarsi di eventi anomali.

Riferimenti bibliografici

[Barracchia *et al.*, 2017] Emanuele Pio Barracchia, Gianvito Pio, Donato Malerba, e Michelangelo Ceci. Identifying lncRNA-Disease Relationships via Heterogeneous Clustering. In *NFMCP 2017, Workshop of ECML-PKDD 2017*, pages 35–48, 2017.

[Ceci *et al.*, 2015] Michelangelo Ceci, Gianvito Pio, Vladimir Kuzmanovski, e Sašo Dŝeroski. Semi-supervised multi-view learning for gene network reconstruction. *PLOS ONE*, 10(12):1–27, 12 2015.

[Korb e Nicholson, 2010] Kevin B. Korb e Ann E. Nicholson. *Bayesian Artificial Intelligence, Second Edition*. CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2010.

[Mignone e Pio, 2018] Paolo Mignone e Gianvito Pio. Positive unlabeled link prediction via transfer learning for gene network reconstruction. In *ISMIS 2018*, pages 13–23, 2018.

[Pio *et al.*, 2013] Gianvito Pio, Michelangelo Ceci, Domenica D'Elia, Corrado Loglisci, e Donato Malerba. A novel biclustering algorithm for the discovery of meaningful biological correlations between micrornas and their target genes. *BMC Bioinformatics*, 14(S-7):S8, 2013.

[Pio *et al.*, 2014] Gianvito Pio, Donato Malerba, Domenica D'Elia, e Michelangelo Ceci. Integrating microrna target predictions for the discovery of gene regulatory networks: a semi-supervised ensemble learning approach. *BMC Bioinformatics*, 15(S-1):S4, 2014.

[Pio *et al.*, 2015] Gianvito Pio, Michelangelo Ceci, Donato Malerba, e Domenica D'Elia. Comirnet: a web-based system for the analysis of mirna-gene regulatory networks. *BMC Bioinformatics*, 16(S-9):S7, 2015.

[Pio *et al.*, 2017] Gianvito Pio, Michelangelo Ceci, Francesca Prisciandaro, e Donato Malerba. LOCANDA: exploiting causality in the reconstruction of gene regulatory networks. In *Discovery Science 2017*, pages 283–297, 2017.

[Pio *et al.*, 2018] Gianvito Pio, Francesco Serafino, Donato Malerba, e Michelangelo Ceci. Multi-type clustering and classification from heterogeneous networks. *Information Sciences*, 425:107–126, 2018.

[Reis-Filho, 2009] Jorge S. Reis-Filho. Next-generation sequencing. *Breast Cancer Research*, 11(3):S12, 2009.