Analisi di dati multi-omici per la predizione della prognosi di pazienti oncologici

Alessandro Beranti - 977702

Febbraio 2023

Durante il periodo di tesi ho lavorato all'interno di AnacletoLAB, laboratorio focalizzato sull'applicazione di metodi di Intelligenza Artificiale su Biologia Molecolare e Medicina facente parte del Dipartimento di Informatica dell'Università degli Studi di Milano. I geni BRCA1 e BRCA2 sono importanti per la salute dei tessuti della mammella e dell'ovaio. Questi geni sono normalmente responsabili della produzione delle proteine che proteggono dalle mutazioni del DNA e prevengono la crescita di cellule anormali. Tuttavia, quando i geni BRCA1 e BRCA2 sono mutati, le proteine che producono non funzionano correttamente e c'è un rischio maggiore di sviluppare tumori al seno e all'ovaio. Le mutazioni dei geni BRCA1 e BRCA2 sono anche associate a un aumento del rischio di tumori della prostata, del pancreas e dello stomaco.

Il lavoro ha riguardato l'uso di tecniche di apprendimento supervisionato, in particolare usando come algoritmo *random forest* [1], per predire la prognosi di un paziente al carcinoma mammario invasivo usando dati provenienti dal gene BRCA. I dati a disposizione sono stati presi dal *Cancer Genome Atlas* [2], programma scientifico di riferimento per la genomica del cancro. I dati si riferiscono alla versione del genoma umano hg19, versione pubblicata nel 2009 che contiene informazioni su sequenze di DNA umano, sequenze di geni, siti di regolazione genica e regioni non codificanti di DNA. I dati sono composti da 4 *dataset* distinti e comprendono: valori di espressione di proteine (proteins), RNA messaggero (mRNA), microRNA (miRNA) e le varianti nel numero di copie genetiche di DNA (CNV). La presenza o assenza di un evento tumorale nel paziente so-

no invece state ottenute grazie all'utilizzo di un *dataset* curato manualmente, noto come TCGA-CDR [3].

In particolare si è voluto verificare se l'utilizzo di dati multi-omici [4] riescano ad aumentare la precisione nella previsione della prognosi. Questo tipo di dati sono un insieme che contiene le variazioni molecolari su più livelli quali: genomica, epigenomica, trascrittomica, proteomica, metabolomica e microbiotica. Dopo una prima fase di *preprocessing*, sono state utilizzate diverse tecniche di riduzione della dimensionalità sia prese singolarmente sia come concatenazione di più tecniche. Lo scopo è stato, in entrambe le soluzioni, cercare di semplificare il *dataset* pur mantenendo contemporaneamente le informazioni più importanti e significative. Successivamente sono stati eseguiti gli esperimenti utilizzando le diverse configurazioni di tecniche di riduzione della dimensionalità applicate ai singoli *dataset* e a ulteriori *dataset* ottenuti mediante la concatenazione degli stessi al fine di cercare la miglior configurazione possibile. Per evitare di sovrastimare la capacità del modello di generalizzare su nuovi dati è stata usata una *Stratified 10-fold cross-validation* in modo da non utilizzare lo stesso insieme di dati per addestrare e valutare il modello. Questa tecnica è stata anche utilizzata all'interno del *tuning* degli iperparametri. Il *dataset* fornito era sbilanciato così si è scelto di utilizzare il valore dell'area sotto la curva *precision-recall* (AUPRC) sia come metrica per valutare la bontà di generalizzazione del modello sia per fare il *tuning* degli iperparametri poiché è una metrica robusta per *dataset* così composti.

Per eseguire questo studio sono state utilizzate diverse tecnologie, prima tra tutte il linguaggio usato è stato *Python* 3.10.6, unitamente a diversi pacchetti specifici per l'analisi dei dati quali: *pandas* e *numpy*. Pacchetti per applicare le tecniche di riduzione della dimensionalità quali: *scipy*, *umap*, *minepy*, *pymrmr* e *boruta*. Infine *sklearn* per l'applicazione specifica di tecniche di apprendimento automatico. Come ambiente di lavoro è stato usato *Jupyter Notebook*, mentre come tool di *versioning*, *git* unitamente a *github*.

Utilizzando come metrica l'AUPRC, la maggior parte degli esperimenti effettuati ha avuto un risultato di previsione della malattia superiore rispetto a una diagnosi casuale, in particolare la seguente concatenazione di tecniche: Pearson + calcolo dimensionalità intrinseca + umap raccoglie i risultati migliori su tutti i *dataset* forniti e creati mediante concatenazione. In particolare usando il *dataset* miRNA è stato possibile classificare come sani tutti i pazienti effettivamente sani (550 su 550) e su 77 pazienti malati il classificatore ne ha riconosciuti come tali 54. Non c'è però evidenza concreta che l'utilizzo

di dati multi-omici riesca ad aumentare la precisione della prognosi della malattia. Uno dei problemi maggiori è stata la composizione del *dataset* che era sbilanciato tra pazienti sani e malati, anche se in campo medico questo è relativamente comune. Su 627 osservazioni solo 77 erano di pazienti effettivamente malati. Questa scarsità di osservazioni positive non ha di certo aiutato il modello a riuscire a generalizzare a dovere.

Riferimenti bibliografici

[1] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[2] `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`.

[3] J. L. et al., "An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics," *Cell*, vol. 173, no. 2, pp. 400–416.e11, 2018.

[4] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biology*, vol. 18, p. 83, May 2017.