UNIVERSITÀ DEGLI STUDI DI TRENTO

Dipartimento di Ingegneria e Scienza dell'Informazione

Corso di Laurea in
Informatica

Elaborato finale

Titolo

Sottotitolo (alcune volte lungo - opzionale)

Supervisore
Prof. Alberto Montresor

Laureando
Giovanni De Toni

Anno accademico 2016/2017

Ringraziamenti

...thanks to...

Indice

Sommario

L'influenza è una infezione respiratoria acuta causata principalmente da virus che possono essere divisi in due ceppi, la varietà A e la varietà B. E' una patologia stagionale, che si presenta spesso durante i mesi invernali (nelle zone con clima temperato) ed è presente in tutto il mondo. Il vettore di trasmissione principale consiste nelle goccioline di muco e saliva, contenenti il virus, che vengono prodotte quando una persona infetta starnutisce o tossisce. Questo la rende una malattia che si diffonde facilmente e rapidamente, specialmente nel caso di zone molto affollate.I sintomi riscontrati spesso sono: febbre alta, tosse, emicrania, dolori articolari e malessere generale. Normalmente, l'infezione sparisce nel giro di una settimana, senza dover ricorrere a particolari cure mediche. In certe categorie a rischio però, se contratta, l'influenza può degenerare e portare anche alla morte [1].

Soltanto in Europa, il Centro Europeo per il controllo delle malattie (ECDC) indica come l'influenza stagionale causi da 4 ai 50 milioni di ammalati e circa 15000-70000 morti annuali a causa dell'infezione [2]. Globalmente, il numero di decessi a causa dell'influenza è di circa da 250000 a 500000 persone all'anno [1]. In Italia l'influenza colpisce mediamente ogni anno l'8% della popolazione [3]. Le categorie più colpite sono sopratutto le fasce di popolazione in età pediatrica (0-4 anni e 5-14 anni) con un incidenza cumulativa che descresce con l'aumentare dell'età. I casi severi e le complicanze sono più frequenti nei soggetti al di sopra dei 65 anni di età, oppure con condizioni di rischio ad esempio malattie cardiovascolari, respiratorie o immunitarie [3].

Essendo una patologia che può colpire la maggior parte della popolazione, essa è considerata un grave rischio per la sanità pubblica e per la collettività. Si stima che nei paesi industrializzati, epidemie influenzali possono generare alti livelli di assenteismo, sia lavorativo che scolastico, e una riduzione della produttività [1]. In Italia, uno studio ha stimato il costo totale delle epidemie influenzali nel periodo 1999-2008 cha varierebbe da 15 a 20 miliardi di euro [4]. La durata media di assenza dal posto di lavoro a causa dell'influenza è di circa 4,8 giorni. Inoltre, i costi diretti in media sono di circa 330 euro a persona (visite, diagnostica, farmaci) e possono salire a circa 3000-6000 euro in caso di ricovero ospedaliero. I costi sociali indiretti (inattività scolastica o lavorativa) ammontano invece a 1000 euro a persona [5].

Attualmente, l'attività dei virus influenzali viene monitorata da alcuni centri facenti parte del Global Influenza Surveillance and Response System (GISRS), un network di sorveglianza globale sponsorizzato dall'WHO. Questi centri forniscono: informazioni sugli attuali ceppi circolanti, indicazioni per la produzione dei vaccini antiinfluenzali (su quali varietà focalizzarsi) ed esaminano e conservano campioni dei virus per scopi di ricerca [6].

Per quanto riguarda l'Italia, il Centro di Controllo delle Malattie (CCM) del Ministero della Salute sostiene che un componente fondamentale per il controllo dell'influenza (sia epidemica che pandemica) è la sorveglianza. Nel nostro paese esistono già programmi di monitoraggio dei livelli di ILI (Influenza-Like Illness), come InfluNet. Influnet è il sistema nazionale di sorveglianza epidemiologica e virologica; il suo compito è quello di stimare l'incidenza settimanale della sindrome influenzale (avvalendosi di dati raccolti da medici sentinella disseminati su tutto il territorio nazionale), in modo da rilevare la durata e l'intensità dell'epidemia [7]. Durante la stagione invernale vengono pubblicati anche dei bollettini settimanali, tramite il servizio FluNews, che illustrano l'evoluzione della situazione italiana [8]. Questi dati vengono anche condivisi sia con il WHO che con ECDC.

InfluNet fornisce informazioni molto importanti riguardo all'incidenza delle patologie influenzali sulla popolazione italiana, però i dati vengono spesso pubblicati con molto ritardo (in media 2 settimane) rispetto all'arco di tempo che descrivono. Per attuare azioni efficaci e coordinare la ditribuzione di

materiale sanitario, produzione di vaccini etc. è necessario avere immediatamente a portata di mano dei dati sulla situazione.

Ci sono stati diversi sforzi per tentare di prevedere o stimare i livelli di incidenza di ILI all'interno della popolazione, sfruttando fonti di dati non covenzionali (cioè non direttamente le informazioni mediche che, come sappiamo, arrivano con un certo ritardo) [9, 10, 11, 12]. Normalmente, i dati che vengono sfruttati maggiormente sono quelli prodotti dai social media, ad esempio: messaggi di Twitter [12], page view di Wikipedia [9, 10] e keywords di ricerca di Google [11]. Da questi studi emerge che attraverso l'utilizzo di tecniche di machine learning è possibile arrivare a delle stime dei livelli di ILI all'interno della popolazione, con settimane di anticipo rispetto ai metodi tradizionali.

L'obbiettivo di questa tesi è cercare di replicare, per quanto possibile, le tecniche dei lavori precedentemente citati, per controllare se anche nella nostra penisola sia possibile effettuare, attraverso tecniche di machine learning, un'analisi attiva per la sorveglianza della diffusione di malattie influenzali. Lo studio più promettente riguarda l'utilizzo delle page view di Wikipedia per stimare il numero di malati settimanali negli Stati Uniti [9], esso quindi verrà utilizzato come base per invece il presente lavoro.

Il Capitolo 1 fornisce una descrizione più dettagliata dei dati che sono stati utilizzati in questo progetto (page view di Wikipedia e bollettini di InfluNet). Inoltre, si definiscono anche i metodi usati per l'analisi degli stessi e alcune informazioni statistiche sulla composizione del dataset. Nel Capitolo 2 vengono descritti i metodi di machine learning che sono stati selezionati per procedere alla creazione del modello predittivo finale. Il Capitolo 3 e il Capitolo 4 presentano rispettivamente: i risultati dell'esperimento e le riflessioni finali su quello che gli esperimenti hanno evidenziato.

Metodi

Analisi dei dati

Wikipedia

Wikipedia (*https://wikipedia.com*) è un enciclopedia libera online, a contenuto libero e gratuito, lanciata da Jimmy Wales e Lerry Sangers nel 2001 ed ora gestita dalla Wikimedia Foundation. Al suo interno sono presenti 45 milioni di voci in oltre 280 lingue. Per quanto rigurda questo lavoro, è stata utilizzata la versione italiana di Wikipedia (*http:// it.wikipedia.org*), che al momento (maggio 2017) conta circa 1400000 articoli [13]. La versione italiana di Wikipedia nel 2016 ha avuto una media giornaliera di 17 milioni di visite [14].

I dati utilizzati i dati delle page view di Wikipedia [15]. Con page view si intende il numero di visite che sono state effettuate su una determinata pagina di Wikipedia in un certo lasso di tempo. I dati disponibili coprono un arco di tempo che va dal Dicembre 2007 al Maggio 2016. Essi sono dati grezzi, nel senso che non distinguono tra visite effettuate da utenti umani oppure tra visite effettuate da bot. Inoltre, essi sono comprensivi delle visite effettuate tramite dispositivi desktop, quindi le visite effettuate da cellulare non sono presenti all'interno delle statistiche.

Influnet

I dati riguardanti i livelli di ILI in Italia sono stati ottenuti attraverso l'analisi dei bollettini Influnet che vengono pubblicati settimanalmente dal Ministero della Salute per tutta la durata del periodo influenzale.

Analisi dei dati

Conclusioni

Bibliografia

[1] Who - influenza (seasonal). http://www.who.int/mediacentre/factsheets/fs211/en/, November 2016. ultimo accesso 22/08/2017.

[2] Ecdc - factsheet about seasonal influenza. https://ecdc.europa.eu/en/seasonal-influenza/facts/factsheet, 2017. ultimo accesso 26/08/2017.

[3] Epicentro - aspetti epidemiologici. http://www.epicentro.iss.it/problemi/influenza/epidItalia.asp. ultimo accesso 22/08/2017.

[4] PL Lai et al. Burden of the 1999-2008 seasonal influenza epidemics in italy: Comparison with the h1n1v (a/california/07/09) pandemic. *Hum Vaccin 7 Suppl.*, pages 217–225, 2011.

[5] Aurelio Sessa, C Lucioni, Gaetano D'Ambrosio, and Germano Bettoncelli. Economic evaluation of clinical influenza in italy. 7:14–20, 01 2005.

[6] Who - influenza, surveillance and monitoring. http://www.who.int/influenza/surveillance_monitoring/en, 2016. ultimo accesso 25/08/2017.

[7] Influnet. http://www.iss.it/iflu, 2013. ultimo accesso 25/08/2017.

[8] Flunews - rapporto epidemiologico settimanale. http://www.epicentro.iss.it/problemi/influenza/FluNews.asp, 2017. ultimo accesso 25/08/2017.

[9] David J. McIver and John S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLOS Computational Biology*, 10(4):1–8, 04 2014.

[10] Kyle S. Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M. Hyman, Alina Deshpande, and Sara Y. Del Valle. Forecasting the 2013–2014 influenza season using wikipedia. *PLOS Computational Biology*, 11(5):1–29, 05 2015.

[11] Google flu trends. https://www.google.org/flutrends/about/, 2014. ultimo accesso 26/08/2017.

[12] Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLOS ONE*, 6(5):1–10, 05 2011.

[13] Wikipedia statistics - charts italian. https://stats.wikimedia.org/EN/ChartsWikipediaIT.htm, 2017. ultimo accesso 26/08/2017.

[14] Analisi visualizzazioni totali it.wikipedia.org. https://tools.wmflabs.org/siteviews/?platform=all-access&source=pageviews&agent=user&range=last-year&sites=it.wikipedia.org, 2017. ultimo accesso 26/08/2017.

[15] Page view statistics for wikimedia projects. https://dumps.wikimedia.org/other/pagecounts-raw/, 2017. ultimo accesso 26/08/2017.