Homework 1

Davide Molitierno

La tesi più importante portata avanti da Andrew NG all'interno della sua intervista per IEEE Specrtum è quella per cui nei sistemi che richiedono l'utilizzo di un determinato data set, nello specifico quelli di IA, in cui è richiesto l'addestramento di un modello attraverso dei dati, ci si debba concentrare su un approccio che sia differente da quello di tipo model centric, portato avanti negli ultimi 15 anni, che consiste nello scaricare un data set già costruito e concentrare i propri sforzi sul migliorare l'architettura del sistema. L'approccio portato avanti da NG è invece quello di tipo data centric per cui si prova a ingegnerizzare i dati effettuando delle operazioni che permettono di migliorare la qualità del data set utilizzato. Nello specifico in presenza di un sottoinsieme di dati non buoni, la pratica precedente prevedeva di arricchire il data set tramite ulteriori dati e successivamente lasciare che fosse l'algoritmo a mediare tra tutti. Nell'approccio data centric piuttosto si cerca di utilizzare tool che etichettano la qualità dei dati in modo tale da poi poter migliorare la consistenza di quello specifico sottoinsieme. Tale processo, detto data cleaning, è fondamentale in IA ma solitamente per la maggior parte dei task viene fatto manualmente. L'obiettivo invece deve essere quello di utilizzare tool che individuano in maniera automatica i dati rumorosi e cercare di concentrarsi su di essi.

A mio parere tale tesi risulta estremamente valida dal momento che l'approccio utilizzato precedentemente ha portato a delle architetture ormai quasi non più migliorabili e che risultano essere estremamente efficaci in quasi qualsiasi contesto. Inoltre molti problemi non necessitano dell'utilizzo di una grande mole di dati ma di data set molto più piccoli e molte aziende non dispongono di giganti data sets a cui attingere, per cui risulta fondamentale incentrare la propria attenzione non più sui big data ma sui good data. Altro valido motivo è quello per cui spesso collezionare più dati di quanti già non se ne posseggono risulta essere un'operazione molto più costosa e talvolta onerosa rispetto ad andare a lavorare sulla qualità del dataset già presente. Infatti è difficile migliorare le prestazioni di un sistema che ha già buone prestazioni per tutto il data set tranne che per un certo sottoinsieme semplicemente cambiando l'intera architettura. Infine un altro importante aspetto introdotto da NG è anche quello dei così detti "synthetic data", tool fondamentale che permette di concentrarsi sul migliorare le prestazioni per una certa classe andando a generare più dati per quella specifica categoria ma, così come NG, ritengo che ci siano prima ulteriori tool più utili da poter utilizzare come migliorare la consistenza nell'assegnamento delle etichette oppure collezionare più dati utili per quella specifica classe.