Abstract

Con l'esplosione dei Big Data, le aziende hanno l'opportunità di accedere a un'enorme quantità di dati che possono migliorare la loro efficienza in termini di processo decisionale, soluzioni adottate, assistenza clienti e così via. Strutturando opportunamente i processi di Estrazione della Conoscenza, le aziende possono convertire facilmente le informazioni in opportunità. Tuttavia, in contesti in continua evoluzione, un'analisi significativa dovrebbe essere dedicata alla valutazione della qualità dei dati per trattare informazioni inaffidabili. Inoltre, le soluzioni decisionali progettate dovrebbero essere consapevoli della deriva dei dati e (ri)adattarsi lungo il loro ciclo di vita.

In questo senso, il lavoro di tesi propone metodologie di Data Mining che tengono conto delle sfide di Veracity e Value alla base dei Big Data. Il significato di Veracity nel contesto dei Big Data riguarda la veridicità di un set di dati e quanto siano affidabili l'origine, il tipo e l'elaborazione dei dati.

Tuttavia, il Valore dei Big Data è strettamente correlato alla Veridicità (o qualità) dei dati trattati. In effetti, la consapevolezza dell'integrità dei dati e delle loro fonti è fondamentale se si cerca di estrarre informazioni da enormi quantità di dati.

Alcuni dei principali risultati di questo lavoro di tesi sono riassunti di seguito:

- L'applicazione della nota teoria della Formal Concept Analysis e delle sue varianti per l'estrazione di modelli di concettualizzazione da diversi contenuti di flussi di dati (es. social media, paper, ecc.).
- La definizione e la sperimentazione di un metodo per mettere in relazione tra loro le fonti di dati, con diversi livelli di velocità, dimensione e credibilità, unendo modelli di concettualizzazione per supportare l'affidabilità delle informazioni (es. Veracity) e abilitare un sistema di filtraggio delle informazioni.
- La definizione e la sperimentazione di un modello di concept-drift aware deep learning basato su LSTM per riconoscere e distinguere in modo adattivo i comportamenti di consumo energetico in evoluzione, eliminando il rischio di falsi positivi sulle frodi.
- La definizione di una misura di consistenza basata sul modello Fuzzy Consensus, un metodo ampiamente utilizzato nel Group Decision Making, per supportare la valutazione del valore dei dati di addestramento prima di applicare un algoritmo di machine learning per l'estrazione di un modello predittivo.

Le metodologie presentate sono supportate dall'applicazione e dalla sperimentazione su diversi scenari applicativi del mondo reale che danno un'idea della loro applicabilità ed efficacia. I problemi affrontati includono raccomandazioni, rilevamento di anomalie, rilevamento di notizie false, farmacovigilanza, sovraffollamento del Pronto Soccorso, ecc.