OR3 - Big Data Analytics

Descrizione sintetica dell'obiettivo

Il risultato raggiunto in questo OR è stata la definizione di algoritmi e tecniche per l'analisi di grosse moli di dati sanitari eterogenei. In particolare sono state i) analizzate tecniche e strumenti di Big Data Analytics (BDA), ii) definiti strumenti per la generazione di statistiche avanzate e per la costruzione di modelli previsionali ed inferenziali e iii) proposti modelli di data mining e Deep Learning per l'elaborazione di dati e l'individuazione di relazioni di causalità non note. Infine, sono state definite metodologie per la costruzione di metriche di correlazione tra parametri/sintomi/contesto clinico e geografico.

L'OR in esame è stato ottenuto con successo attraverso le attività seguenti.

A3.1 – Analisi di tecniche e di strumenti innovativi di Big Data Analytics

Obiettivo della Attività

In questa attività saranno analizzate nel dettaglio le tecniche e le metodologie innovative per la gestione delle problematiche relative ai Big Data Analytics.

Data la grossa quantità di dati da elaborare e l'eterogeneità delle fonti, l'analisi e la gestione delle informazioni saranno svolte con nuovi approcci basati sulla BDA, che nasce proprio per risolvere tali problemi. Essa può essere vista come l'unione di due concetti: i Big Data e la Business Analytics.

Nello specifico, sono stati dimostrati in letteratura gli enormi vantaggi che la Sanità può trarre dall'applicazione delle tecniche di BDA; la crescente disponibilità di una sempre più grande mole di dati in ambito sanitario ha evidenziato la necessità anche in tale contesto di soluzioni performanti, evolute e dedicate, che siano capaci di gestire la varietà dei dati, sia provenienti da fonti certificate che non, e l'esigenza di trasformare questi ultimi in conoscenza il più velocemente possibile.

Descrizione Attività

Nell'ambito dell'attività A3.1 i risultati raggiunti sono stati la valutazione e la scelta di framework per la gestione di grosse moli di dati sanitari. A tal fine, sono state analizzate nel dettaglio le problematiche legate al fenomeno dei "Big Data", ovvero la capacità di immagazzinare, gestire e analizzare grandi quantità di dati eterogenei. Inoltre è stata fatta una review degli strumenti che meglio si prestano a questo scopo. In letteratura sono stati dimostrati gli enormi vantaggi che la sanità può trarre dall'applicazione delle tecniche di BDA; la crescente disponibilità di dati in ambito sanitario ha evidenziato la necessità anche in tale contesto di soluzioni performanti, evolute e dedicate, che siano capaci di gestire l'eterogeneità dei dati, sia provenienti da fonti certificate che non, e l'esigenza di trasformare questi ultimi in conoscenza il più velocemente possibile. Tra i numerosi strumenti presenti nel contesto Big Data, è emersa la necessità di individuare framework che potessero garantire elevate prestazioni, scalabilità e disponibilità di librerie. Sono stati effettuati test comparativi in termini di velocità di elaborazione, occupazione CPU, memoria e dischi, banda di rete richiesta e scalabilità. A valle delle analisi e test condotti si è optato per la scelta dei framework Apache Spark, MongoDB, Apache Hadoop e Spark SQL.

A3.2 – Data integration and aggregation

Obiettivo della Attività

Questa attività ha lo scopo di definire tecniche avanzate per il pre-processing di grosse moli di dati sanitari eterogenei, allo scopo di favorirne l'integrazione, la correttezza e la trasformazione in aggregati da cui poter derivare informazioni consistenti e prive di errori. I dati, una volta acquisiti dalla piattaforma di cloud computing, necessitano infatti di una fase preliminare di data aggregation e mash-up, permettendo di ottenere soluzioni rapide ed affidabili per la loro

memorizzazione in forma strutturata ed il loro allineamento, favorendo l'elaborazione delle interrogazioni e l'estrazione di informazioni.

Saranno definiti algoritmi e metodologie di MapReduce per definire tecniche avanzate di ETL (Extraction, Transformation and Loading) per gestire questa fase allo scopo di effettuare il processo di riconciliazione, in termini di integrazione, pulizia e trasformazione in dati espressi in un formato integrato ed uniforme da cui poter derivare informazioni consistenti e prive di errori, rispettando anche i limiti temporali imposti dall'applicazione.

Descrizione Attività

Nell'attività A3.2 i risultati ottenuti sono stati la definizione di tecniche e metodi di pulizia e integrazione del dato che meglio si adattano al contesto Big Data, necessari a causa degli elevati tassi di incongruenze, errori ed eterogeneità presenti nelle grandi moli di dati da elaborare. A tal fine, sono state definite tecniche avanzate per il pre-processing di dati sanitari eterogenei, allo scopo di favorire l'integrazione, la correttezza e la trasformazione in aggregati da cui poter derivare informazioni consistenti e prive di errori. I dati, una volta che sono stati acquisiti dalla piattaforma, necessitano di una fase preliminare di data aggregation e mash-up, permettendo di ottenere soluzioni rapide ed affidabili per la loro memorizzazione in forma strutturata ed il loro allineamento, favorendo l'elaborazione delle interrogazioni e l'estrazione di informazioni. Sono stati valutati algoritmi e metodologie di MapReduce per definire tecniche avanzate di ETL (Extraction, Transformation and Loading) per effettuare il processo di riconciliazione, in termini di integrazione, pulizia e trasformazione dei dati espressi in un formato integrato ed uniforme da cui poter derivare informazioni consistenti e prive di errori, rispettando anche i limiti temporali imposti dall'applicazione. Inoltre, sono state analizzate, oltre a MapReduce, altre tecniche di ETL specifiche per Big Data, in particolare Parallel ETL e Streaming ETL.

A3.3 – Definizione di tecniche per la generazione di statistiche avanzate

Obiettivo della Attività

In questa attività saranno analizzate e progettate tecniche statistiche avanzate, basate su sistemi BDA, per la realizzazione di report avanzati che consentano di rappresentare graficamente le informazioni desiderate, fornendo differenti viste all'interno delle quali le misure di interesse potranno essere riferite a varie dimensioni o poste in dipendenza di esse.

Nella realizzazione dell'attività si intende integrare un approccio che adatti le tecniche e le metodologie per l'elaborazione ed analisi di Big Data con le metodologie per fornire supporto nei processi decisionali, legati in particolare alla diagnosi corretta e tempestiva, alla selezione dei trattamenti, alla prognosi, all'assistenza. In maggior dettaglio, l'aggregazione e l'analisi dei dati saranno finalizzate alla produzione di evidenze per consentire la valutazione di diverse alternative e le loro implicazioni a livello personale e sociale e le modalità attraverso cui le macro-informazioni di carattere generale possano essere utili per generare dei bollettini statistici.

Descrizione Attività

I risultati raggiunti nella attività A3.3 sono culminati nell'analisi delle tecniche per la generazione di statistiche. A tal fine sono state approfondite le tecniche per la generazione di statistiche avanzate ottenute da Big Data e per la realizzazione di report complessi, che consentono di rappresentare graficamente le informazioni desiderate fornendo differenti viste relative alle misure di interesse. L'attività si è focalizzata sulla definizione di metodologie che hanno permesso di elaborare i dati clinici e sanitari aggregati, permettendo di ottenere rappresentazioni grafiche multidimensionali. In particolare, le tecniche hanno permesso di: i) analizzare la totalità delle informazioni in conformità a dimensioni differenti e le aggregazioni trasversali (slicing), ii) focalizzare ed estrarre una sezione avente particolare interesse (dicing), iii) scomporre i dati nelle sue determinanti all'interno della stessa gerarchia (drill-down), iv) scomporre il problema nelle sue determinanti passando da una gerarchia ad un'altra (drill-across), v) aggregare i dati, vi) rappresentare il problema attraverso differenti approcci, quali: Basic Analytics per Insights, Analisi Fattoriale, Analisi delle corrispondenze binarie e multiple.

A3.4 – Modelli previsionali dei dati, metriche di correlazione e algoritmi di data mining

Obiettivo della Attività

In questa attività saranno analizzate tecniche di data mining con lo scopo di fornire strumenti per l'identificazione di informazioni nascoste all'interno di grandi volumi di dati, con particolare riferimento a metodologie statistiche e matematiche quali alberi decisionali, cluster analysis e regole di associazione.

Sulla base dei dati elaborati ed archiviati, saranno definite tecniche e strumenti che consentano di definire correlazioni e inferenze tra fattori critici di successo e indicatori di performance a supporto dell'analisi dei processi. In particolare, l'applicazione di questa metodologia avrà l'obiettivo di favorire il raggiungimento di molteplici obiettivi strategici spesso in conflitto tra loro. Grazie alla disponibilità dei dati eterogenei acquisiti dal sistema e alle metodologie di aggregazione sviluppate ad hoc, sarà possibile definire metodi inferenziali su dati relativi ai soggetti, a gruppi di popolazione, ad aree geografiche, atti a evidenziare correlazioni (farmacologiche, patologiche, ambientali, ecc.), oltre a ottenere dati e modelli epidemiologici.

Descrizione Attività

Nell'attività A3.4 sono state valutate tecniche per realizzare previsioni e classificazioni sui dati. Lo studio ha fatto riferimento a metodologie statistiche e matematiche quali alberi decisionali, cluster analysis e regole di associazione. Sulla base dei dati elaborati ed archiviati, sono state delineate le tecniche e gli strumenti necessari per definire le correlazioni e le inferenze tra fattori critici di successo e gli indicatori di performance a supporto dell'analisi dei processi. Grazie alla disponibilità dei dati eterogenei acquisiti dal sistema e alle metodologie di aggregazione sviluppate ad hoc, è stato possibile definire metodi inferenziali su dati relativi ai soggetti, a gruppi di popolazione, ad aree geografiche, atti a evidenziare correlazioni (farmacologiche, patologiche, ambientali, ecc.), oltre a ottenere dati e modelli epidemiologici. Inoltre, sono state definite le metriche per consentire di analizzare le ricadute in termini di benefici, salute e fattori di rischio medico derivanti dall'utilizzo delle nuove tecnologie da parte del singolo e della collettività. Grande rilevanza è stata data alle tecniche di Intelligenza Artificiale, le quali, attraverso l'analisi di grosse quantità di informazioni, sono in grado di fornire previsioni e realizzare analisi predittive (predictive analytics), fornendo ulteriori informazioni, con vari livelli di dettaglio come, ad esempio, abitudini dei soggetti, analisi per cluster o tipologia, sino alle classiche informazioni legate ad aree geografiche, periodicità, categorie di individui, andamenti stagionali e così via. Gli approcci usati sono stati sia di tipo Deep Learning (reti convoluzionali e ricorsive e auto-encoders) che di tipo Machine Learning (Support Vector Machines e Clustering).