Big Data

ACCESSIBILI AGLI UTENTI APPASSIONATI

Un contributo del Politecnico di Milano

Laura M. Sangalli – Letizia Tanca

Big Data e Data Science, attualmente parole chiave nel mondo del business e dell'industria, stanno dando vita a uno dei più ferventi campi della ricerca, con sinergiche interazioni tra la statistica, la matematica e l'informatica. Molteplici sono le linee di ricerca di frontiera elaborate in questo ambito dal Politecnico di Milano. Ne illustreremo due, che tentano di rispondere alla grande sfida che la Data Science deve affrontare: lo sviluppo di sistemi per l'esplorazione di grandi masse d'informazioni e di tecniche sofisticate di analisi di dati massivi sempre più complessi.

Questo contributo illustra alcune delle linee di ricerca nell'ambito Big Data sviluppate al Politecnico di Milano. L'analisi di grandi quantità di dati, spesso molto complessi e ricchi dal punto di vista semantico, come nel caso di oggetti che hanno numerosi attributi (o dimensioni), ha ricevuto attenzione negli ultimi anni in quasi tutti i settori della conoscenza. Nella prima parte si illustra una ricerca che ha lo scopo di mettere anche gli utenti non esperti in condizione di comprendere i fenomeni e prendere decisioni basate sui dati, fornendo un sistema che li guidi interagendo come in un dialogo innescato da semplici interrogazioni. Nella seconda parte si affronta l'altra grande sfida della Data Science: l'analisi di dati che, oltre a essere sempre più massivi, sono sempre più articolati, presentando una varietà di aspetti diversificati. Quindi, si espongono alcune tecniche informatiche e statistiche per l'analisi di dati compositi ad alta dimensionalità.

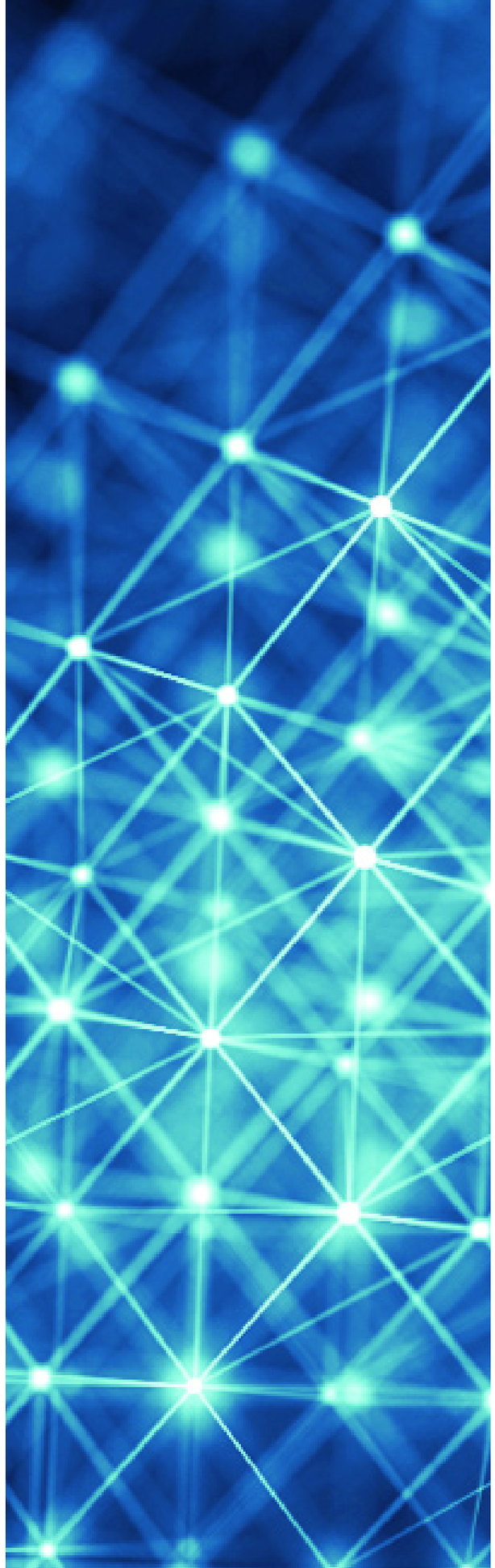
EXPLORATORY COMPUTING: L'APPROCCIO ESPLORATIVO A GRANDI DATASET

I dati sono disponibili come mai prima e creano un tesoro d'informazioni che sembra in attesa del momento giusto per essere utilizzato. Al Politecnico di Milano, in collaborazione con l'Università della Basilicata, lavoriamo a un approccio chiamato Exploratory Computing (Ec), il cui scopo è riuscire a evidenziare a un utente, che non necessariamente abbia una preparazione tecnica o statistica, gli aspetti rile-

vanti e degni di nota di un dataset troppo grande e ricco di aspetti diversi per essere letto per intero. L'Ec è basato sull'idea che anche gli utenti non esperti (che non siano informatici o analisti di dati) possano essere posti nelle condizioni di eseguire interrogazioni (*query*) complesse su un certo insieme di dati, ad esempio per comprendere un fenomeno, per assumere decisioni ecc., ottenendo risposte sintetiche che li orientino. L'idea è che un utente 'entusiasta' possa dialogare col sistema come con un altro essere umano, dove la persona che chiede non desidera come risposta una lista di record quanto una sintesi ragionata. Supponiamo che un uomo d'affari abbia interesse a investire in Polonia e consulti un esperto di quel paese. Alla domanda: «Come funzionano in Polonia gli aspetti fiscali per un'azienda appena fondata?» non si attenderà l'elenco delle leggi fiscali polacche né una tabella di aliquote, bensì una breve sintesi delle differenze con quel che accade per le aziende italiane. E per fare questo, il sistema deve essere in grado di valutare quali siano gli aspetti 'rilevanti' negli adempimenti fiscali polacchi rispetto a quelli italiani. È in questi termini che si parla di «esplorazione»: una sessione di dialogo così impostata permette all'utente di esplorare velocemente i contenuti rilevanti del dataset d'interesse. D'altra parte, perché si parla di «computing»? Perché, per supportare un'interazione nella quale l'utente persegua gli obiettivi di cui sopra, passando magari da una domanda all'altra in breve tempo, è necessario avere a disposizione grandi capacità di calcolo.

Visti dalla prospettiva Big Data, i problemi presentati da Ec sono numerosi e stimolanti:

– VELOCITÀ E INTEGRAZIONE DELLE TECNICHE. Nella comunità di ricerca delle basi di dati, i metodi per estrarre da questi conoscenza significativa e sintetica sono perlopiù utilizzati indipendentemente l'uno dall'altro; ad esempio, per risolvere un problema specifico si può adottare una tecnica di Data Mining oppure di statistica. Invece, lo scopo di Ec è di mettere a disposizione di utenti inesperti un sofisticato sistema che applichi di volta in volta la tecnica più adatta all'ispezione e configuri il risultato come un percorso di ricognizione dove, a ogni passo, il sistema presenti agli utenti dei fatti rilevanti e concisi sui dati, per permettergli di comprendere e progredire verso il passo di esplorazione successivo;
– LA NOZIONE DI RILEVANZA È FONDAMENTALE, poiché il compito principale del sistema è richiamare l'attenzione dell'utente sulle interessanti (o sorprendenti) differenze o somiglianze tra gli insiemi di dati. Supponiamo che una persona interroghi il sistema chiedendo informazioni sui musei del Sud Italia. Piuttosto che fornire l'elenco di nomi e indirizzi, un sistema di Ec risponderà con una serie di caratteristiche, come il fatto



che i musei del Sud, rispetto agli altri nel paese, contengano un numero superiore di reperti greci e romani, e che quelli greci siano eseguiti con metalli meno preziosi o, addirittura, che i musei del Sud costino mediamente meno di quelli del Nord. Questo tipo di risposte richiede che il sistema sia in grado di decidere cosa è rilevante nell'insieme dei musei del Sud rispetto a tutti gli altri. L'identificazione di sottoinsiemi interessanti, ottenuta combinando i valori degli attributi – e cioè le caratteristiche dei dati – può prendere strade molto interessanti. Per esempio, il sistema potrebbe 'imparare' gli interessi dell'utente e presentare non gli aspetti significativi in assoluto ma quelli che lo sono specificamente per lui. Un altro approccio potrebbe essere quello di rivelare all'utente delle proprietà di nicchia, vale a dire quelle che hanno senso solo in piccoli sottoinsiemi del dataset ecc;
– ACCENTO SUL CONCETTO D'INTENSIONALITÀ (in senso logico). Le risposte cosiddette intensionali non forniscono elenchi di oggetti ma li descrivono in base alle loro proprietà. A un motore di ricerca, una query sugli hotel di Milano fornirà all'utente un elenco di alberghi; la stessa domanda, rivolta a un amico, probabilmente darà luogo a considerazioni come «la maggior parte degli hotel sono piuttosto costosi (riferimento a tutto il set) ma quelli intorno alla stazione centrale sono più economici e comunque decenti (riferimento a un sottoinsieme creato dinamicamente), proprio come quelli vicini all'università, ma questi ultimi sono più rumorosi». Queste sono caratteristiche intensionali, che permettono all'utente di elaborare un'idea sintetica e soddisfacente di ciò che gli interessa;
– INTERAZIONE BASATA SUL DIALOGO. Imitando il comportamento dell'essere umano, che in un dialogo risponde all'interlocutore facendo riferimento a quello che ha detto e 'costruendoci' sopra, Ec mira a supportare interazioni dove l'utente può costruire sui feedback forniti in precedenza dal sistema senza doverlo inizializzare ogni volta da zero e facendo liberamente riferimento alle domande e alle risposte precedenti.

I possibili domini applicativi e gli scenari dell'Exploratory Computing sono tutti quelli che si prestano a una conversazione durante la quale un interlocutore cerca di imparare qualcosa da una persona più esperta. Gli esempi sono innumerevoli: noi l'abbiamo applicata al caso dell'esplorazione di dati relativi a passati disastri ambientali, a scopo preventivo – per capire le correlazioni tra i comportamenti umani (costruzioni, demolizioni ecc.) e l'impatto che questi hanno

durante una catastrofe – e anche a supporto della Protezione civile, per stabilire buone prassi d'intervento alla luce di esperienze già vissute. Una seconda applicazione è l'esplorazione di dati medici dai quali estrarre informazioni sintetiche su trend o su prassi consolidate; ad esempio, in risposta a una semplice ricerca dei pazienti i cui test di funzionalità tiroidea sono alterati, il sistema potrebbe rilevare una particolare distribuzione dei valori di età.
Il calcolo della distribuzione dei dati è solo uno dei tanti modi per valutare la rilevanza muovendo da descrizioni concise; si possono impiegare anche altre misure, come l'entropia, che stabilisce il 'livello di varietà' di un insieme di valori, o descrivere i dati mediante modelli di Data Mining. Con tecniche di Data Mining si potrebbe scoprire una relazione frequente tra i valori degli attributi, come il fatto che il 60% dei pazienti con valori di funzionalità tiroidea alterati vivono o hanno vissuto in una località di mare. Come si vede, le tecniche esistenti di statistica, di Data Mining o di machine learning in generale possono essere impiegate con profitto a supporto dell'Exploratory Computing che, dal punto di vista degli strumenti di analisi, non si propone di fare di più. L'accento della ricerca è invece centrato sulla capacità del sistema di fornire, in tempi accettabili, risposte intelligibili per un utente non tecnico, che siano rilevanti ai fini della comprensione e agiscano stimolando ulteriormente la curiosità.

L'analisi di dati complessi e ad alta dimensionalità

Un'altra grande sfida che la Data Science deve affrontare è l'analisi di dati, che oltre a essere sempre più massivi, si configurano via via più strutturati. La loro complessità, unita all'alta dimensionalità, richiede lo sviluppo di tecniche sofisticate per l'esplorazione e la riduzione dimensionale. Riportiamo qui due illustrazioni: una nell'ambito delle neuroscienze e l'altra relativa allo studio della mobilità sulla base di dati di telefonia mobile (figura 1).

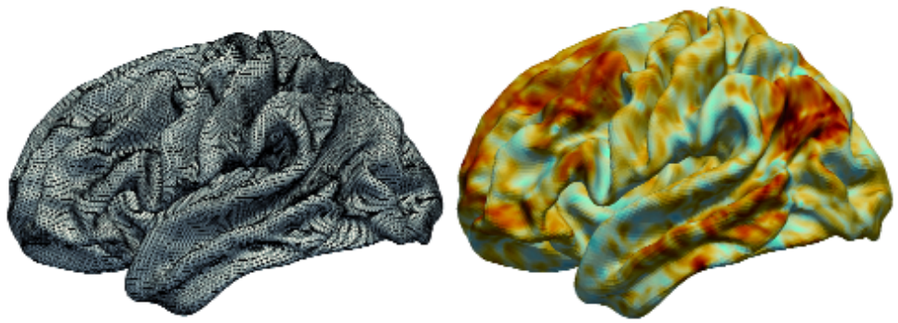


Figura 1. Studio della connettività cerebrale sulla base di segnali altamente dimensionali da neuro-imaging. A sinistra, superficie che approssima l'emisfero sinistro della corteccia cerebrale; a destra, mappa di connettività ottenuta da segnale di risonanza magnetica funzionale.

Un affascinante esempio di dati complessi e ad alta dimensionalità è offerto dalle neuroscienze. In particolare, il recente forte aumento di dati derivanti da neuro-imaging ha portato a profondi cambiamenti nella nostra comprensione del cervello e dei processi cerebrali. La neuroscienza è un campo multidisciplinare e il ruolo dell'analisi dei dati è fondamentale al suo successo. Gran parte del lavoro svolto è consistito nello stabilire come utilizzare i modelli statistici per l'analisi delle grandi masse di dati che derivano da modalità di imaging, quali la risonanza magnetica, l'elettroencefalografia e la magnetoencefalografia. È indubbia la necessità di incorporare una sempre più composita informazione sulla struttura e l'anatomia del cervello nell'analisi statistica, per migliorare la nostra attuale comprensione dei processi cerebrali. Si considerino, ad esempio, i segnali registrati mediante risonanza magnetica funzionale (fMRI), che rilevano un livello di ossigeno del sangue come una sequenza di misurazioni ripetute nel tempo, producendo una serie storica d'immagini tridimensionali. Una maggiore attività neurale in una particolare zona del cervello provoca un aumento della domanda di ossigeno. Poiché il segnale fMRI è riconducibile a variazioni di concentrazione di deossi-emoglobina, questo è considerato una misura surrogata dell'attività neurale e viene utilizzato per produrre mappe di attivazione o per indagare la connettività funzionale del cervello. Gran parte di questo segnale è relativo alla corteccia cerebrale, un sottile foglio di tessuto neuronale che costituisce la parte più esterna del cervello, caratterizzato da una morfologia altamente convoluta; la maggior parte dell'attività neurale è, infatti, focalizzata su di essa. L'anatomia della corteccia può essere estratta da risonanza magnetica strutturale, una tecnica di scansione non invasiva applicata per visualizzare la struttura interna del cervello, rendendolo come immagine tridimensionale ad alta risoluzione spaziale. È naturale rappresentare la corteccia cerebrale come una superficie inclusa in uno spazio tridimensionale e strutturata con una distanza geodetica in due dimensioni piuttosto che con la distanza euclidea all'interno del volume. Aree funzionalmente distinte, che sono distanti lungo la corteccia, potrebbero risultare vicine se misurate con la distanza euclidea nel volume, per via della morfologia particolarmente convoluta della corteccia. Per questo motivo, quando si analizzano i segnali distribuiti su corteccia, trascurarne la morfologia può portare a stime non accurate.
Di recente è stato stabilito come sia vantaggioso analizzare i dati neuro-imaging utilizzando tecniche vincolate alla superficie, per le quali sono state sviluppate alcune metodologie che permettono di analizzare il segnale fMRI concernente un singolo soggetto, che può essere poi mappato su un modello comune di superficie corticale, un cosiddetto atlante, per consentire l'analisi statistica multisoggetto.
Al Politecnico di Milano, in collaborazione con l'Università di Cambridge, nel Regno Unito, abbiamo proposto una prima tecnica che permette di compiere un'analisi multisoggetto di segnali fMRI, considerando l'anatomia della corteccia. L'idea s'ispira a metodi statistici avanzati di analisi delle componenti principali regolarizzate e sfrutta tecniche di analisi numerica, quali gli elementi finiti superficiali. Il metodo favorisce l'esecuzione di una riduzione dimensionale del dato, identificando i pattern principali di connettività comuni

ai vari soggetti e riesce inoltre a gestire efficientemente l'enorme massa dei dati, relativa ai segnali fMRI di molteplici soggetti, grazie a tecniche di programmazione avanzata.
Attualmente vi è un forte impulso nella comunità internazionale per lo sviluppo di metodi per l'analisi di questi dati, e grandi database di fMRI e risonanze magnetiche strutturali sono disponibili al pubblico. Il consorzio Human Connectome Project gestisce un archivio pubblico di scansioni strutturali di fMRI in stato di riposo e in attività, eseguite su un gran numero di volontari, il cui studio è fondamentale per l'avanzamento della conoscenza sul funzionamento del cervello e per la comprensione dei meccanismi basilari di alcune patologie cerebrali.

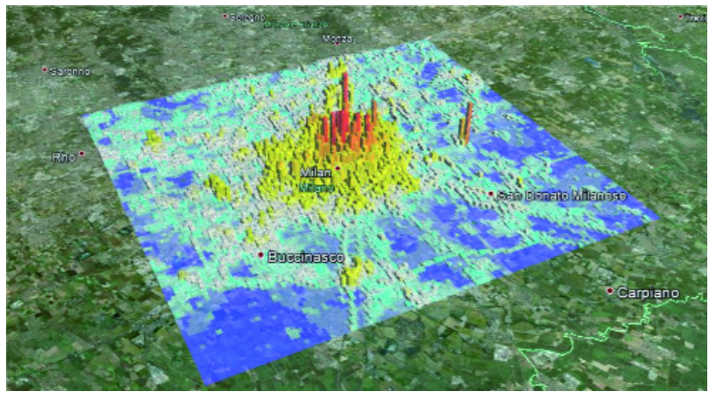


Figura 2. Studio della mobilità sull'area metropolitana milanese sulla base di dati di telefonia mobile (utilizzo medio delle rete mobile nel periodo 18-31 marzo 2009).

Un altro esempio di dati massivi e complessi è offerto dal *Green Move* – un progetto di mobilità sostenibile sviluppato dal Politecnico di Milano e cofinanziato da Regione Lombardia – con l'obiettivo di ideare e sperimentare un nuovo sistema di car sharing con veicoli elettrici per l'area milanese. Per questo sono stati studiati dati di telefonia mobile, resi disponibili a seguito di una convenzione tra Telecom Italia e il Politecnico, e in particolare Telecom Italia ha fornito il dato Erlang, ovvero il numero medio di telefoni cellulari connessi alla rete per chiamate, misurato ogni 15 minuti e riferito a un reticolo uniforme di più di 10.000 celle rettangolari di circa 200 x 300 metri, che copre l'intera area metropolitana milanese. In prima approssimazione, questa quantità può essere considerata proporzionale al numero di persone attive in ogni cella del reticolo in un determinato momento, fornendo informazioni sulla mobilità e sulle dinamiche di popolazione. Oltre alla dimensionalità dei dati (un milione di record al giorno), la struttura del sottostante tessuto urbano determina una forte articolazione risultando, ad esempio, fortemente influenzata dalla rete stradale, con le principali arterie di scorrimento veloce, e dalla rete dei trasporti pubblici, così come dalla diversa destinazione delle varie aree urbane, dai quartieri residenziali ai distretti commerciali e lavorativi, dai grandi impianti sportivi alle strutture fieristiche.

Questi dati sono di forte interesse non solo per la compagnia telefonica stessa, che deve poter prevedere e gestire possibili sovraccarichi locali della rete, ma anche per altri privati, quali i gestori di sistemi di vehicle sharing. Inoltre, il dato è di interesse per la gestione dei servizi d'ordine pubblico e sanitario, per un'efficiente dislocazione delle risorse sul territorio. In un'ottica di medio-lungo periodo, le informazioni estraibili da tali dati sono fondamentali per una pianificazione urbanistica che consenta di diminuire, ad esempio, la congestione che penalizza l'area urbana, con ricadute apprezzabili sull'economia e sulla qualità della vita della metropoli.
Presso il Politecnico di Milano, con uno sforzo congiunto di vari Dipartimenti, sono state sviluppate tecniche avanzate di riduzione dimensionale di questi dati, che permettono di esprimere l'intero segnale, in ordine al traffico di telefonia mobile, in vari sottosegnali, identificando determinati pattern, localizzati nel tempo e nello spazio. Tra questi, alcuni sono ricorrenti, come il sottosegnale derivante dal pendolarismo o quello relativo alla residenzialità, e altri sono anomali, quali i sottosegnali associati a determinati grandi eventi sportivi o fieristici. La Data Science è una disciplina in fortissima crescita e in rapida evoluzione. Gli sforzi descritti nell'affrontare l'esplorazione e l'analisi di grandi masse di dati complessi sono solo alcune prime risposte alle problematiche economiche, in un campo stimolato ogni giorno da nuove sfide.

BIBLIOGRAFIA

M. Buoncristiano et al., *Database challenges for exploratory computing*, «ACM SIGMOD Record» 44 (2015) 2, pp. 17-22.
M. K. Chung et al., *Cortical thickness analysis in autism with heat kernel smoothing*, «NeuroImage» (2005) 25, pp. 1256-1265.
B. Ettinger – S. Perotto – L.M. Sangalli, *Spatial regression models over two-dimensional Manifolds*, «Biometrika» 102 (2016) 1, pp. 71-88.
E. Lila – J.A.D. Aston – L.M. Sangalli, *Smooth Principal Component Analysis over two-dimensional manifolds with an application to neuroimaging*, «Annals of Applied Statistics» 10 (2016) 4, pp. 1854-1879.
F. Manfredini et al., *Treelet Decomposition of Mobile Phone Data for Deriving City Usage and Mobility Pattern in the Milan Urban Region*, in Paganoni – Secchi (eds.) 2014.
M. Mazuran – E. Quintarelli – L. Tanca, *Data Mining for* XML *Query-Answering Support*, «IEEE Transactions on Data and Knowledge Engineering» 24 (2012) 8, pp. 1393-1407.
K. Morton et al., *Support the Data Enthusiast: Challenges for Next-Generation Data-Analysis Systems*, «PVLDB – Proceedings of the Very Large DataDase Conference» 7 (2014) 6, pp. 453-456.
A. Paganoni – P. Secchi (eds.), *Advances in Complex Data Modeling and Computational Methods in Statistics*, Springer, Milano 2014.
P. Paolini – N. Di Blas, *Exploratory portals: The need for a new generation*, International Conference on Data Science and Advanced Analytics (Dsaa), Shanghai 2014, pp. 581-586.
P. Secchi – S. Vantini – V. Vitelli, *Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan*, «Statistical Methods and Applications» 24 (2015) 2, pp. 279-300.
P.-N. Tan – M. Steinbach – V. Kumar, *Introduction to Data Mining*, Pearson, Harlow 2006.