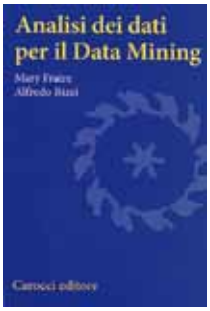
RECENSIONI



M. Fraire e A. Rizzi
Analisi dei dati per il data mining
Carocci Editore, Roma, 2011
pp. 414 € 41,00
ISBN 978-88-430-6033-7

Il libro parte dell'esigenza, presente da qualche decennio, di esaminare una enorme quantità di dati che sono a disposizione della statistica e di molte altre discipline (fisica, biologia, etc.). Il libro è scritto per gli statistici tenendo presenti le esigenze delle Industrie e della Pubblica Amministrazione. L'ampio respiro culturale e tecnico fa ritenere che la lettura del testo sia interessante anche a studiosi di altre discipline come per esempio della fisica. Infatti soprattutto nella fisica delle particelle elementari con l'aumentare della grandezza e della complessità degli apparati sperimentali la quantità dei dati prodotti risulta sempre più elevata e implica un uso massiccio sia dell'hardware che del software comparabile con le esigenze di Istituti come l'ISTAT.

Per illustrare l'ampio respiro culturale del libro si cita la parte finale della prefazione:

"Sotto il profilo concettuale, logico filosofico, nell'analisi dei dati è accentuata l'impronta empirica, la cui validità non è pertanto vincolata a quella di modelli formalizzati o di teorie generali. Gli schemi di riferimento sono quasi sempre di tipo descrittivo. Ci si allontana, in questa prospettiva, dalle impostazioni del filosofo viennese Karl Popper, secondo cui lo scienziato tende a trovare una spiegazione dei fatti osservati. Una descrizione degli stessi deve essere deducibile dalla teoria in congiunzione con le cosiddette *condizioni iniziali*. Si è più vicini, invece, alle posizioni di Francesco Bacone, secondo il quale un esperimento determinante può verificare o rafforzare una teoria; Popper invece afferma che esso può al più confutarla o *falsificarla*. La tradizione della scuola statistica italiana si innesta sul principio di considerare tutti gli aspetti della nostra disciplina in termini di *fasi della conoscenza*, per passare dal dato alla formulazione e, quindi, alla verifica delle teorie."

Il testo è diviso in tre parti oltre a delle utilissime appendici.

Nella prima parte, 'le basi dell'analisi dei dati', si ha una esposizione completa delle tecniche matematiche usate in statistica accompagnate da esempi tratti dalle esigenze quotidiane delle Regioni, dello Stato e delle Aziende. Vengono inoltre messe in evidenza *le fasi dell'analisi dei dati* (AMD) in sette fasi ciascuna della quali viene corredata da esempi concreti e utili alla comprensione.

Nella seconda parte, 'le tecniche dell'analisi dei dati', vengono evidenziati i metodi di classificazione in particolare l'analisi dei gruppi o 'cluster analysis' dove si evidenzia che è un metodo tipicamente esplorativo in cui sono essenzialmente assenti gli aspetti inferenziali. Per chiarire meglio:

"I problemi di classificazione si sono presentati nella scienza sin dalle origini. Infatti la fantasia del ricercatore non è sufficiente per dare origine al processo di formazione di teorie scientifiche e di ideazione di modelli interpretativi della realtà che ci circonda, sia nell'ambito delle scienze naturali sia in quello delle scienze sociali: tutto poggia sulla classificazione dei dati disponibili alla quale si riconduce il momento deduttivo di formulazione di ipotesi di gran parte della ricerca scientifica."

"In alcuni campi di ricerca si può pertanto ritenere che la fase di classificazione sia il momento essenziale del procedimento scientifico; ciò vale anche nei casi in cui una teoria sia intuita prescindendo da prove, come non infrequentemente avviene nel campo della fisica e, talora, nel campo delle scienze sociali."

"Nonostante i progressi raggiunti nel campo della velocità di elaborazione e dalla teoria degli algoritmi, alcuni metodi di CA, come vedremo, sono applicabili solo per meno di un centinaio di oggetti. Nelle situazioni concrete di ricerca che interessano le aziende, si riescono a classificare anche milioni di unità con algoritmi adeguati che, però, forniscono partizioni non aventi tutte le caratteristiche di ottimalità."

Nella terza parte, 'metodi di seconda generazione', vengono trattate le reti neurali, gli alberi di decisione e infine il Data Mining.

Per precisare meglio il Data Mining:

"Il data mining (miniera di dati, to mine è una espressione inglese utilizzata nel senso di scavare per estrarre nelle miniere) è un processo non elementare di evidenziazione di relazioni, dipendenze, correlazioni, associazioni, modelli, strutture, tendenze, classi, fattori ottenuti navigando in grandi insiemi di dati generalmente residenti su banche dati. La navigazione avviene con metodi matematici, statistici, informatici o algoritmici. Questo processo può essere iterativo e/o interattivo a seconda degli obiettivi da raggiungere. Anche se non è contenuto esplicitamente nella definizione, il DM è un processo (il più automatico possibile) che va dai dati elementari disponibili in un warehouse alla decisione con un apporto in ciascun passo di un valore aggiunto informazionale che può portare a decisioni operative in funzione dell'informazione di sintesi evidenziata. Dietro il concetto di DM vi è l'eredità dell'intelligenza artificiale e dei sistemi esperti."

"Per fare DM occorrono competenze differenziate di:
- informatica;
- algoritmica;
- tecnologia dell'informazione;
- analisi dei dati;
- conoscenze aziendali

che possono esprimersi solo in gruppi di lavoro interdisciplinari.

Riveste, inoltre, grandissima importanza la conoscenza, alle volte parziale, del processo di decisione dei dirigenti [delle istituzioni o delle industrie o aziende]."

Infatti: "Il Data Mining richiede l'integrazione dei risultati dell'analisi con processi decisionali".

Da segnalare nelle appendici una trattazione della rilevazione degli ascolti televisivi (Auditel) e una valutazione critica del software statistico, che peraltro è presente anche in altri punti del testo.

In conclusione si evidenzia che l'alto livello del libro ne consiglia la lettura soprattutto in relazione alla parte matematica. Inoltre si auspica un reciproco scambio tra il mondo della fisica e quello della statistica ai fini di un confronto delle tecniche.

R. Habel, M. Pallotta
INFN - Laboratori Nazionali di Frascati