Introduzione al KDD e al DATA MINING

Vincenzo Antonio Manganaro

vincenzomang@virgilio.it, www.statistica.too.it

Indice

1 Verso il DM: una breve analisi delle fasi del processo KDD.

Gli stadi che caratterizzano un processo KDD sono stati identificati nel 1996 da Usama Fayyad, Piatetsky-Shapiro e Smyth (Fig 1.1). Nell'elencare e descrivere queste fasi tali studiosi pongono particolare attenzione allo stadio del DM, cioè a tutti quegli algoritmi per l'esplorazione e lo studio dei dati. Il DM è ritenuta la fase più importante dell'intero processo KDD e questa sua enorme importanza, peraltro riconosciuta, rende sempre più difficile, soprattutto in termini pratici, distinguere il processo KDD dal DM. Da parte nostra, anche se alcuni ricercatori usano i termini

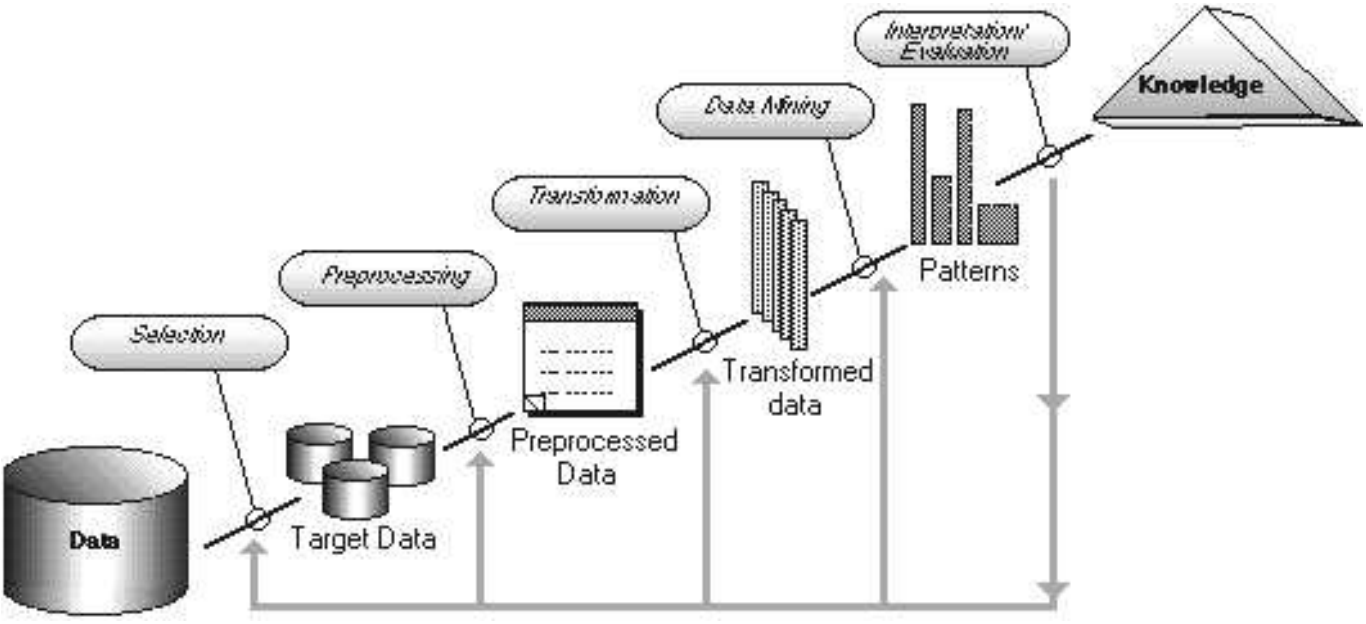

Figura 1: Fasi del processo KDD secondo Fayyad, Piatetsky-Shapiro e Smyth.

DM e KDD come sinonimi, cercheremo costantemente di separare i due aspetti e di considerare il DM la fase più significativa del processo KDD, ma non perfettamente coincidente con esso. Iniziamo con una descrizione delle cinque fasi del processo KDD che ricalca sostanzialmente il modello ideato dagli studiosi citati sopra.
Il processo KDD prevede come dati in input dati grezzi e fornisce come output informazioni utili ottenute (fig 1.1) attraverso le fasi di:

- **Selezione**: i dati grezzi vengono segmentati e selezionati secondo alcuni criteri al fine di pervenire ad un sottoinsieme di dati, che rappresentano il nostro *target data* o dati obiettivo. Risulta abbastanza chiaro come un database possa contenere diverse informazioni, che per il problema sotto studio possono risultare inutili; per fare un esempio, se l'obiettivo è lo studio delle associazioni tra i prodotti di una catena di supermercati, non ha senso conservare i dati relativi alla professione dei clienti; è invece assolutamente errato non considerare tale variabile, che potrebbe invece fornire utili informazioni relative al comportamento di determinate fasce di clienti, nel caso in cui si voglia effettuare un'analisi discriminante[4].

[4]L'esempio non è casuale, ma è riferito all'applicazione che considereremo nel capitolo 8. In quella sede si eliminerà dal database di partenza, tra le altre variabili, anche la variabile professione,

- **Preelaborazione**: spesso, pur avendo a disposizione il target data non è conveniente nè, d'altra parte, necessario analizzarne l'intero contenuto; può essere più adeguato prima campionare le tabelle e in seguito esplorare tale campione effettuando in tal modo un'analisi su base campionaria. Fanno inoltre parte del seguente stadio del KDD la fase *di pulizia dei dati (data cleaning)* che prevede l'eliminazione dei possibili errori e la decisione dei meccanismi di comportamento in caso di dati mancanti.

- **Trasformazioni**: effettuata la fase precedente, i dati, per essere utilizzabili, devono essere trasformati. Si possono convertire tipi di dati in altri o definire nuovi dati ottenuti attraverso l'uso di operazioni matematiche e logiche sulle variabili. Inoltre, soprattutto quando i dati provengono da fonti diverse, è necessario effettuare una loro riconfigurazione al fine di garantirne la consistenza.

- **Data Mining**: ai dati trasformati vengono applicate una serie di tecniche in modo da poterne ricavare dell'informazione non banale o scontata, bensì interessante e utile (lo studio analitico di tali tecniche sarà oggetto della seconda parte della tesi). I tipi di dati che si hanno a disposizione e gli obiettivi che si vogliono raggiungere possono dare un'indicazione circa il tipo di metodo/algoritmo da scegliere per la ricerca di informazioni dai dati. Un fatto è certo: l'intero processo KDD è un processo interattivo tra l'utente, il software utilizzato e gli obiettivi, che devono essere costantemente inquadrati, ed iterativo nel senso che la fase di DM può prevedere un'ulteriore trasformazione dei dati originali o un'ulteriore pulizia dei dati, ritornando di fatto alle fasi precedenti.

- **Interpretazioni e Valutazioni**: il DM crea dei pattern, ovvero dei modelli, che possono costituire un valido supporto alle decisioni. Non basta però interpretare i risultati attraverso dei grafici che visualizzano l'output del DM, ma occorre valutare questi modelli e cioè capire in che misura questi possono essere utili. È dunque possibile, alla luce di risultati non perfettamente soddisfacenti, rivedere una o più fasi dell'intero processo KDD.

---

che non ha alcun senso in un contesto di basket analysis, ovvero di studio delle associazioni dei prodotti venduti nei supermercati, oggetto del nostro studio.

2 Il DM: Alcune definizioni.

Non possiamo certo affermare che il concetto di DM sia ben delimitato. Gli sviluppi di tale tecnica, per lo più abbastanza recenti, e gli ampi campi di applicazione, soprattutto nelle diverse tipologie di business, fanno si che il DM risulti spesso un concetto piuttosto vago e caratterizzato da varie definizioni. Ecco allora le definizioni più comuni di DM:

- *Il DM è la non banale estrazione di informazione implicita, precedentemente sconosciuta e potenzialmente utile attraverso l'utilizzo di differenti approcci tecnici* (Frawley, Piatetsky-Shapiro e Matheus, 1991).

- *Il DM è una combinazione di potenti tecniche che aiutano a ridurre i costi e i rischi come anche ad aumentare le entrate estraendo informazione dai dati disponibili* (T.Fahmy).

- *Il DM consiste nell'uso di tecniche statistiche da utilizzare con i databases aziendali per scoprire modelli e relazioni che possono essere impiegati in un contesto di business* (Trajecta lexicon).

- *Il DM è l'esplorazione e l'analisi, attraverso mezzi automatici e semiautomatici, di grosse quantità di dati allo scopo di scoprire modelli e regole significative* (Berry, Linoff, 1997).

- *Il DM è la ricerca di relazioni e modelli globali che sono presenti in grandi database, ma che sono nascosti nell'immenso ammontare di dati, come le relazioni tra i dati dei pazienti e le loro diagnosi mediche. Queste relazioni rappresentano una preziosa conoscenza del database e, se il database è uno specchio fedele, del mondo reale contenuto nel database.* (Holshemier e Siebes, 1994).

- *Il DM si riferisce all'uso di una varietà di tecniche per identificare "pepite" di informazione e di conoscenza per il supporto alla decision making. L'estrazione di tale conoscenza avviene in modo che essa possa essere usata in diverse aree come supporto alle decisioni, previsioni e stime. I dati sono spesso voluminosi ma, così come sono, hanno un basso valore e nessun uso diretto può esserne fatto; è l'informazione nascosta nei dati che è utile* (Clementine user guide).

Una definizione poco formale ma molto intuitiva che dà certamente il senso di cosa è il DM è la seguente:

- *Il DM è una vera e propria tortura dei dati al fine di farli confessare....*

3 Un modello standard per il DM: il CRISP-DM.

Nel paragrafo precedente abbiamo detto che il DM non è ancora un concetto ben delimitato. Tuttavia esiste un progetto finanziato dalla Commissione europea il cui obiettivo è quello di definire un approccio standard ai progetti di DM, chiamato CRISP-DM (CRoss Industry Standard Process for Data Mining). Il CRISP-DM affronta la necessità di tutti gli utenti coinvolti nella diffusione di tecnologie di DM per la soluzione di problemi aziendali. Scopo del progetto è definire e convalidare uno schema d'approccio indipendente dalla tipologia di business[5]. La figura 1.2 riassume lo schema CRISP-DM, oltre che chiarire l'essenza del DM e il suo utilizzo da parte delle imprese per incrementare il loro business.
Come possiamo vedere dalla figura il ciclo di vita di un progetto di DM consiste di sei fasi la cui sequenza non è rigida. È quasi sempre richiesto un ritorno indietro ed un proseguimento tra le differenti fasi. Ciò dipende dalla bontà del risultato di ogni fase, che costituisce la base di partenza della fase successiva. Le frecce indicano le più importanti e frequenti dipendenze tra le fasi. L'ellisse fuori lo schema rappresenta la natura ciclica di un processo di DM il quale continua anche dopo che una soluzione è stata individuata e sperimentata. Spesso quanto imparato durante un processo di DM porta a nuove informazioni in processi di DM consecutivi.
Descriviamo ora con maggiore dettaglio le fasi della figura:

- **Business Understanding:** è opportuno che in un progetto di DM si conosca il settore di affari in cui si opera. In questo senso il DM non deve, nè può sostituire il compito dei manager tradizionali, ma solo porsi come strumento aggiuntivo di supporto alle decisioni. Avendo chiare le idee sul settore di affari in cui si opera, si procede alla conversione di questa conoscenza di settore nella definizione di un problema di DM e quindi alla stesura preliminare di un piano prefissato per raggiungere gli obiettivi stabiliti.

[5]Per maggiori informazioni sul progetto si veda l'URL: http://www.spss.it/datamine/crisp_is.html.

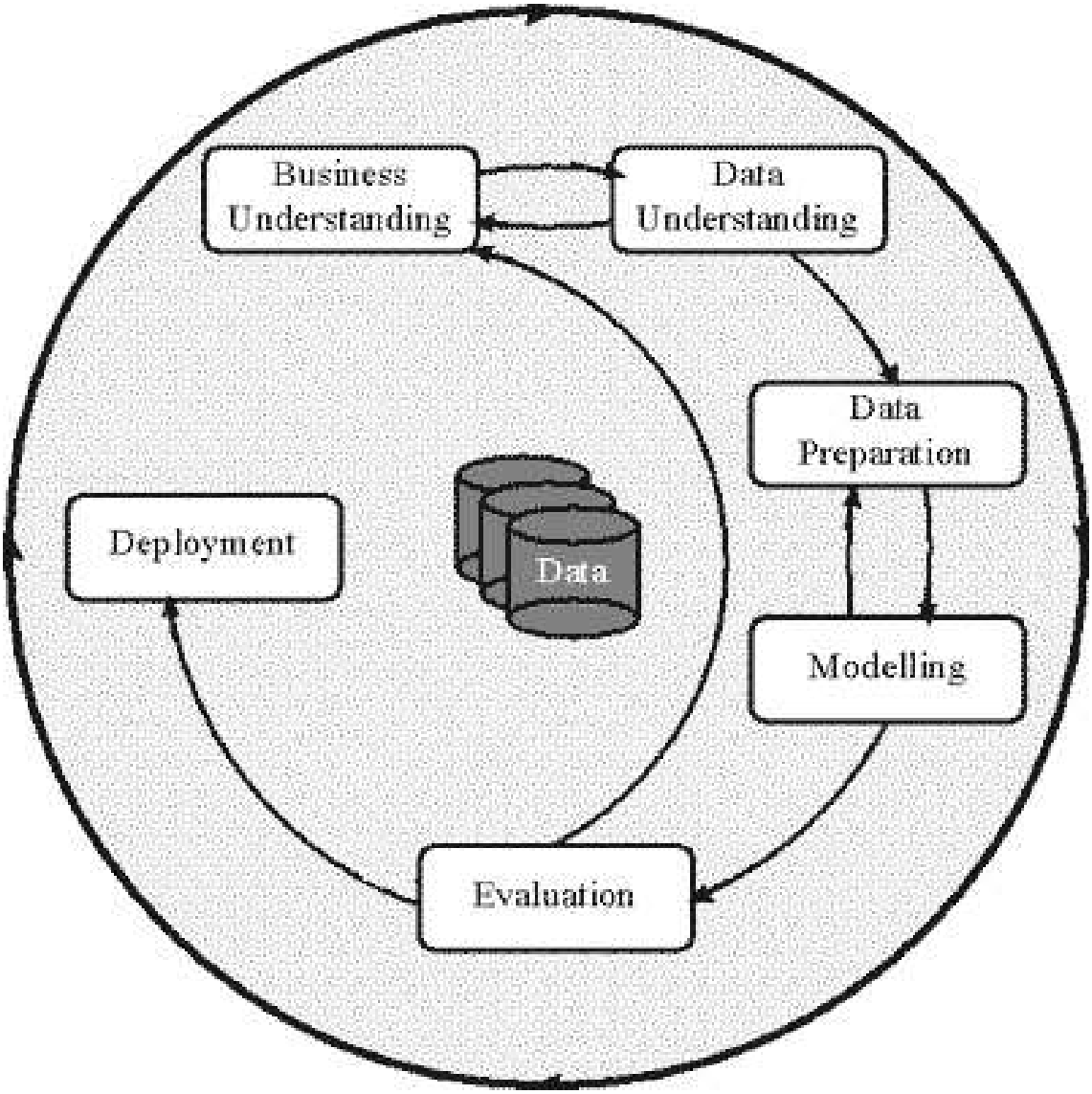

Figura 2: Fasi del CRISP-DM

- **Data Understanding:** individuati gli obiettivi del progetto di DM, ciò di cui disponiamo per il raggiungimento di tali obiettivi è rappresentato dai dati. Quindi la fase successiva prevede una iniziale raccolta dei dati e una serie di operazioni sui dati stessi che permettono di acquisire maggiore familiarità con essi, di identificare problemi nella qualità dei dati stessi, nonché scoprire le prime informazioni che a volte si possono ricavare dal semplice calcolo delle statistiche di base (medie, indici di variabilità, ecc. . .). È chiaro inoltre come le prime due fasi siano collegate dato che rappresentano l'individuazione dei fini e dei mezzi di un progetto di DM.

- **Data preparation:** tale fase copre tutte le attività che poi portano alla costruzione dell'insieme di dati finale a partire dai dati grezzi e dunque dell'insieme di dati cui applicare le tecniche di DM. Essa comprende tra l'altro la selezione di tabelle, di records e di attributi come anche, se necessaria, la trasformazione e la pulitura dei dati.

- **Modelling:** in questa fase vengono selezionate e applicate varie tecniche che permettono di ricavare dei modelli. Determinate tecniche, per poter essere applicate, necessitano di specifiche richieste rispetto alla forma dei dati, per cui è spesso opportuno tornare indietro alla fase di preparazione dei dati per modificare il dataset iniziale e adattarlo alla tecnica specifica che si vuole utilizzare.

- **Evaluation:** prima di procedere all'impiego del modello o dei modelli costruiti, è molto importante valutare il modello e i passi eseguiti per costruirlo, accertarsi che attraverso tale modello si possono veramente raggiungere obiettivi di business, capire se qualcosa di importante non è stato sufficientemente considerato nella costruzione del modello.

- **Deployment:** è la fase finale che prevede l'utilizzo del modello o dei modelli creati e valutati che possono permettere il raggiungimento dei fini desiderati.

Le fasi di DM su descritte sembrano ricalcare nella sostanza le fasi del più generale processo di estrazione di conoscenza dai database (KDD). In realtà questo progetto di DM ingloba al suo interno le fasi del processo KDD e dimostra quanto già affermato in precedenza riguardo al sempre più diffuso accostamento del DM al processo

KDD. D'altro canto uno studio anche di base delle tecniche di DM -come del resto si propone questa tesi- non può assolutamente escludere le fasi che non riguardano la specifica applicazione di tali tecniche. Per questo, soprattutto nella parte applicativa del presente lavoro, svilupperemo anche le fasi del KDD che non riguardano il DM, o secondo questo nuovo approccio, le fasi di un progetto di DM che non riguardano specificamente la fase del Modelling.

4 DM di tipo descrittivo e previsivo: Verification models e Discovery models.

Gli approcci al DM possono essere di due tipi: di tipo *top-down* e di tipo *bottom-up*. Nel primo caso si tratta di utilizzare la statistica come guida per l'esplorazione dei dati, cercando di trovare conferme a fatti che l'utente ipotizza o già conosce, o per migliorare la comprensione di fenomeni parzialmente conosciuti. In quest'ambito vengono utilizzate le statistiche di base, che permettono di ottenere descrizioni brevi e concise del dataset, di evidenziare interessanti e generali proprietà dei dati; è anche possibile l'utilizzo di tecniche statistiche tradizionali come, ad esempio, la regressione. Tuttavia, un approccio di tipo top-down limita i compiti del DM ad un DM di tipo descrittivo .

La sola descrizione dei dati non può fornire quelle informazioni di supporto alle decisioni, cui si fa costantemente riferimento quando si parla di potenzilità del DM. Di conseguenza, un approccio al DM di tipo bottom-up, nel quale l'utente si mette a scavare nei dati alla ricerca di informazioni che apriori ignora, risulta di gran lunga più interessante. Questo secondo approccio conduce ad un DM di tipo previsivo in cui si costruisce uno o più set di modelli, si effettuano delle inferenze sui set di dati disponibili e si tenta di prevedere il comportamento di nuovi dataset. È proprio nel DM di tipo previsivo che IBM ha identificato due tipi, o modi, di operare, che possono essere usati per estrarre informazioni di interesse per l'utente: i **Verification models** e i **Discovery models**.

I verification models utilizzano delle ipotesi formulate dall'utente e verificano tali ipotesi sulla base dei dati disponibili. In questi tipi di modelli riveste un ruolo cruciale l'utente, al quale è affidato il compito di formulare delle ipotesi sui possibili comportamenti delle variabili in questione. Tuttavia risultano di gran lunga più in-

teressanti i discovery models, che costituiscono la parte più rilevante delle tecniche di DM. In questi tipi di modelli all'utente non è affidato nessun tipo di compito specifico, è il sistema che scopre "automaticamente" importanti informazioni nascoste nei dati: si cerca di individuare pattern frequenti, regole di associazione, valori ricorrenti. Potremmo addirittura affermare che il DM è costituito dai soli discovery models, dal momento che un'importante differenza tra i metodi tradizionali di analisi dei dati e i nuovi metodi (di cui il DM è parte integrante) è che i primi sono guidati dalle assunzioni fatte, nel senso che viene formulata un'ipotesi che viene saggiata attraverso i dati, mentre i secondi sono guidati dalla scoperta, nel senso che i modelli sono "automaticamente" estratti dai dati. L'utilizzo di queste nuove tecniche richiede ovviamente enormi sforzi di ricerca e in questo le maggiori performance degli odierni calcolatori giocano un ruolo chiave. Un utilizzo delle tecniche tradizionali di analisi dei dati si ha nelle procedure OLAP (On Line Analytical Processing). Come affermato nella nota 3 dell'introduzione, si tratta di tecniche alternative al DM che a differenza di questo hanno alla base un'analisi essenzialmente deduttiva. Difatti, in tali procedure l'utilizzatore "interroga" il database ponendo una serie di "queries" che vanno a validare o meno un'ipotesi precedentemente formulata. Tuttavia, quando il numero delle variabili cresce, l'utilizzo delle metodologie OLAP diventa sempre più difficoltoso, perché diventa difficile, e anche dispendioso in termini di tempo, formulare delle buone ipotesi da saggiare. Risulta quindi più utile ricorrere alle tecniche di DM che liberano l'utente da compiti specifici, dal momento che in tale ambito non si utilizzano più strumenti di Query e OLAP, ma tecniche derivate dalla statistica e dall'intelligenza artificiale.

5 DM e OLAP: Tecniche alternative o complementari?

Ritengo utile riportare un esempio tratto dal sito internet della società SPSS la quale, fra le altre attività, comprende anche quella della implementazione e diffusione di algoritmi di DM. Nell'esempio la procedura OLAP e il DM vengono viste come tecniche complementari: il DM consente agli utenti di strumenti OLAP di andare oltre i report riassuntivi. Il DM dice **perché** un certo fenomeno sta succedendo, mentre l'OLAP si limita a dire **cosa** sta succedendo. Per esempio, il DM può sco-

prire gruppi di clienti o di prodotti che condividono caratteristiche simili. Per capire cosa significa, diamo uno sguardo ai dati giornalieri di acquisizione di clienti di una banca (grafici 1.3 e 1.4). Il risultato dello strumento OLAP è un grafico (grafico 1.3) che fornisce un'informazione molto chiara: l'acquisizione di clienti sta seguendo un trend positivo, nonostante abbia avuto una flessione nei mesi centrali dell'anno. Se però si procede con ulteriori analisi e si affiancano alle tecniche OLAP le tecniche di DM, emergono delle informazioni diverse. Procedendo ad una segmentazione dei clienti fatta tramite una cluster analysis, si osserva la distribuzione della clientela rappresentata nel grafico 1.4. Il grafico 1.4 mostra che l'acquisizione sta aumentando fra i clienti di Breve termine (B_T), è sostanzialmente stabile fra i clienti definiti Generici (GEN) e sta calando fra quelli di Lungo termine (L_T). Dal momento che i clienti di Lungo termine sono i più interessanti per la banca, questa tendenza rappresenta un problema. Disponendo di questa ripartizione è stato innanzitutto possibile rilevare il problema, e sarà possibile studiare azioni specifiche di marketing dirette a invertire la tendenza. È quindi evidente che gli strumenti OLAP rappresentano una base di partenza, ma non sono in grado di fornire lo stesso contributo informativo delle tecniche di DM. Tuttavia l'esempio dimostra come le tecniche di OLAP e DM siano tecniche complementari piuttosto che alternative.

6 DM: Potenzialità, limiti e campi di applicazione.

Ormai pensiamo di aver sufficientemente chiarito in che cosa consista il DM, dando anche un'idea delle potenzialità di tali strumenti; bisogna comunque fare attenzione a non sopravvalutarli: non è assolutamente detto che l'applicazione di metodi di DM in un contesto aziendale possa risolvere problemi specifici con una certa facilità e con costi contenuti. Il DM è infatti una "tecnica di frontiera", difficilmente esso risolve grossi problemi, piuttosto aiuta ad individuare piccoli particolari, che in un contesto altamente competitivo, quale quello attuale, possono fare la differenza per le organizzazioni che ne fanno uso. Quale che sia il campo di applicazione, il DM non elimina il bisogno di conoscere alla perfezione il settore in cui si opera, di capire i dati che si hanno a disposizione e di capire il funzionamento dei metodi analitici usati; esso può assistere i manager nel trovare modelli e relazioni nei dati, ma questi

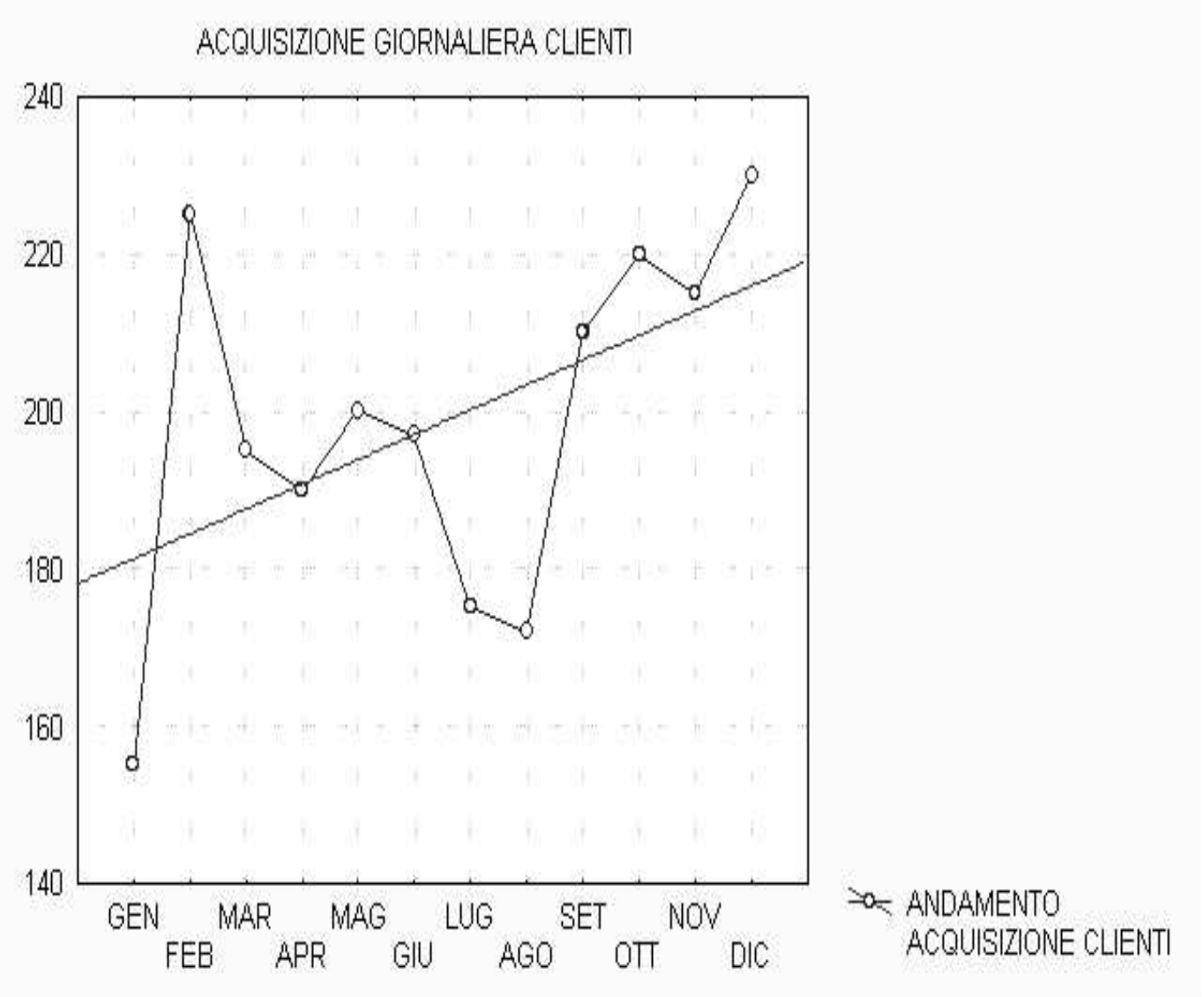

Figura 3: Risultati di una procedura OLAP.

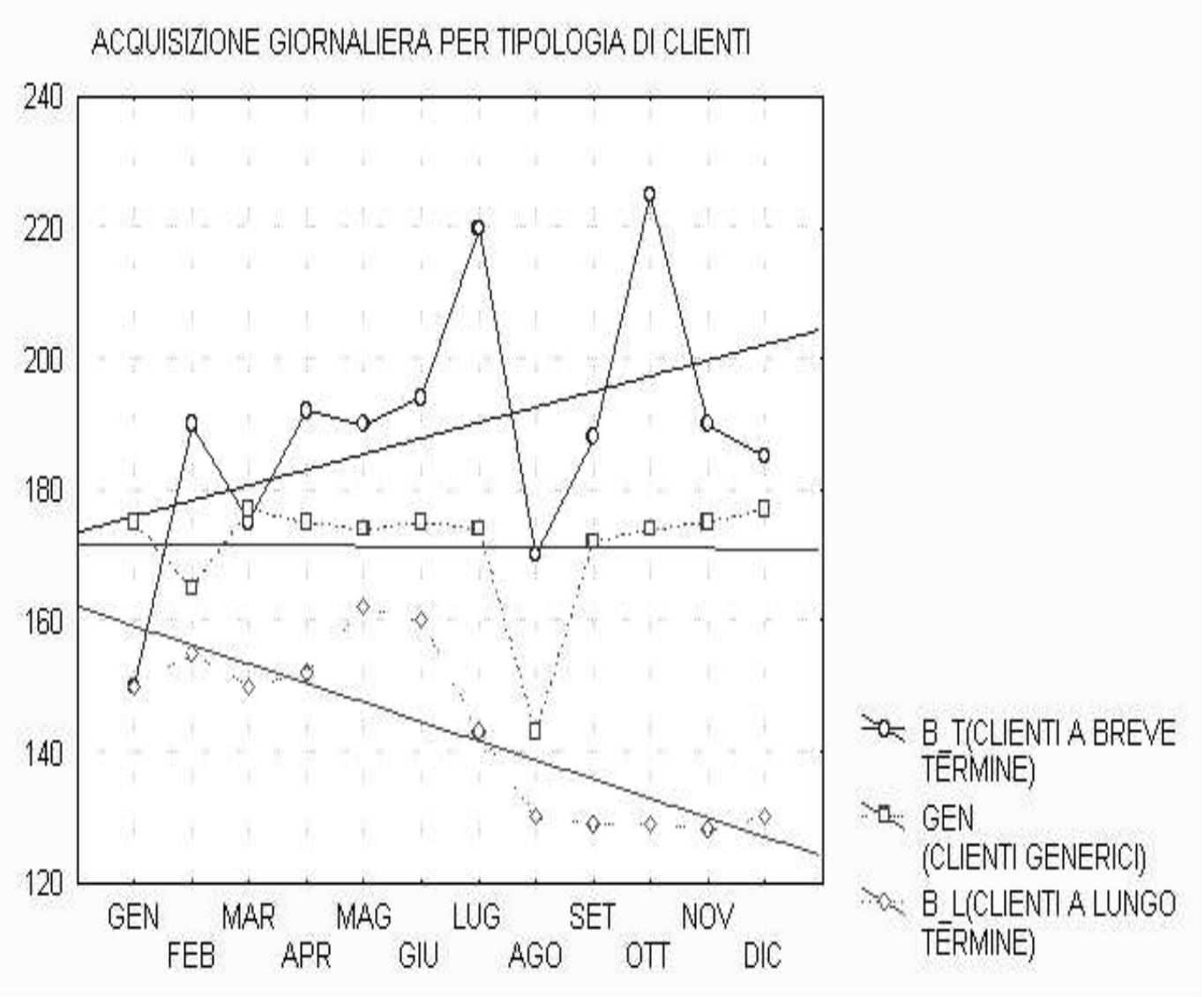

Figura 4: Risultati di una procedura di DM.

modelli devono essere costantemente verificati nel mondo reale. Concludiamo questo capitolo introduttivo dando un'idea di quanto grande possa essere il campo di applicazione del DM e del processo KDD, facendo riferimento ad alcune applicazioni tipiche. Ciò può servire anche da stimolo ad approfondire le conoscenze di tali strumenti, vista l'enorme utilità ed applicabilità che hanno in molti settori aziendali. Alcune applicazioni abbastanza interessanti riguardano il marketing strategico ed in particolare la market basket analysis, attraverso la quale è possibile pianificare determinate campagne promozionali e di cui peraltro considereremo un'applicazione nel capitolo 8. Le applicazioni di DM coinvolgono anche il settore bancario con analisi della vulnerabilità dei clienti (risk analysis), l'individuazione delle truffe (fraud detection) nella telefonia cellulare e nelle carte di credito, ecc....