

Università degli Studi di Modena e Reggio Emilia
Facoltà di Scienze della Comunicazione (Comunicazione, Economia, Informazione)
Corso di Laurea in Economia, Reti, Informazione
(curriculum in Economia delle Reti e della Comunicazione)
Anno Accademico 2003/04 – primo semestre

Analisi dei Dati e Data Mining

prof. Alessandro Zanasi

P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, A. Zanasi
Discovering Data Mining - From Concept to Implementation
(Prentice Hall PTR, Upper Saddle River, New Jersey 07458)

Chapter 1: “Data Mining: the Basics”

Back to the Future

Nel 1956, quello che poteva essere un piccolo negozio di hardware a conduzione familiare, si trovava ad operare in un contesto profondamente diverso da quello attuale. Prima della produzione di massa e dell'avvento dei mass-media, cioè prima del mass marketing, i comportamenti d'acquisto dei consumatori erano stabili, così com'erano stabili i loro modelli di preferenza. I nuovi entranti (potenziali concorrenti) si trovavano in difficoltà, per via del forte grado di intimità che i negozi di più lungo corso avevano raggiunto con i propri clienti.

Tale situazione è mutata profondamente al giorno d'oggi. È infatti scomparso il tradizionale customer relationship ed il comportamento dei consumatori è divenuto molto più volatile che in passato. Un numero sempre maggiore di organizzazioni si sta però rendendo conto del valore aggiunto costituito da una relazione più intima con i propri consumatori.

Why Now?

La maggior parte del recente interesse affluito nel data mining deriva da due forze principali: i drivers (cioè le necessità) e gli enablers (ovvero i mezzi per implementarlo). I drivers sono in primo luogo le modifiche dell'ambiente di business, che ha portato a mercati sempre più competitivi. Tra gli enablers, particolare importanza hanno avuto i più recenti avanzamenti tecnici (ricerche sul machine learning, sui db e sulle tecnologie di visualizzazione).

Changed Business Environment

Modifiche fondamentali stanno influenzando il modo in cui le organizzazioni osservano e pianificano l'approccio con i propri consumatori:

- modelli d'acquisto dei consumatori: i consumatori sono sempre più esigenti ed hanno accesso alle migliori informazioni disponibili in un dato momento. D'altro canto, molti

utenti, in una sorta di crisi di rigetto, stanno limitando il numero di negozi in cui effettuano acquisti. Vi è inoltre un mutamento demografico: solo il 15% delle famiglie statunitensi è di tipo “tradizionale” (un uomo, una donna ed uno o più figli);

- saturazione del mercato: molti mercati sono ormai saturi (ad esempio quelli dei conti corrente, delle automobili e delle assicurazioni). Chi vuole incrementare la propria quota di market share deve quindi ricorrere a politiche di acquisizione/fusione o praticare prezzi predatori;
- nuovi mercati di nicchia: stanno emergendo molti nuovi mercati altamente focalizzati su particolari esigenze dei consumatori;
- aumentata commoditization: è sempre più difficile osservare una reale differenziazione dei beni e dei servizi. Occorre dunque cercare nuovi canali distributivi;
- esaurimento dei tradizionali approcci al marketing: gli approcci del marketing di massa e del database marketing sono sempre più inefficaci. I consumatori ricercano canali più specifici;
- time to market: il time to market di un prodotto, specialmente nel campo dell’information technology, è sempre più importante (vedi Netscape, che con un vantaggio di soli pochi mesi sui rivali, ha guadagnato l’80% del mercato dei browser in meno di un anno);
- ciclo di vita dei prodotti: i prodotti sono venduti sul mercato più rapidamente, ma hanno solitamente un ciclo di vita molto più corto che in passato (nel campo dell’IT è nata la definizione di “web year”, corrispondente a tre mesi);
- aumento della competizione e dei rischi di business: è sempre più difficile tenere traccia di tutte le forze competitive. Inoltre, i rapidi mutamenti dei trends dei consumatori, infondono ulteriori rischi in qualsiasi attività di business.

Drivers

Contro tale background, diverse organizzazioni stanno rivalutando i loro tradizionali approcci al fare business, cercando la via per rispondere ai mutamenti dell’ambiente in cui si trovano ad operare. I drivers che guidano questo processo di rivalutazione sono:

- focus sul consumatore: rinnovamento del customer relationship alla ricerca di maggiore intimità, collaborazione e partnership 1-to-1. Le organizzazioni devono rispondere a nuove domande sui propri consumatori (ad esempio: “quali classi di consumatori ho?”, “esistono sottoclassi con comportamenti d’acquisto simili?”, “come posso vendere di più alla mia clientela abituale?”, ecc...);
- focus sulla concorrenza: le organizzazioni devono aumentare l’attenzione nei confronti delle forze competitive, in modo da poter predisporre un arsenale di armi di business. Ad esempio, occorre predire le future strategie dei principali competitors, i movimenti tattici dei competitors locali, quali sono i propri clienti particolarmente sensibili alle offerte della concorrenza, ecc...;
- focus sul data asset: diverse organizzazioni iniziano ora a vedere il loro archivio di dati accumulati negli anni, come una risorsa critica di business. A spingere in questa direzione sono i migliori indici ROI fatti registrare dalle imprese che adottano, in chiave di decision

making, tecniche guidate dai dati (ad esempio il datawarehousing) e la crescente disponibilità di data warehouses (più l'approccio al datawarehouse si diffonde, più i pionieri investono per migliorare i propri sistemi e mantenere così il vantaggio competitivo precedentemente acquisito).

Enablers

Combinati ai drivers, gli enablers spingono verso una revisione del tradizionale approccio al decision making. Tra gli enablers possiamo individuare:

- data flood: la computerizzazione della vita quotidiana ha fatto sì che organizzazioni di vario genere abbiano a disposizione moltissimi dati relativi ai comportamenti quotidiani degli individui. Questi dati vengono sempre più usati anche per la gestione del business, ad esempio tracciando trends o cercando di scoprire nuove opportunità di business. In più, fonte importante per i data miners, sono oggi disponibili moltissimi dati demografici (modelli di scelta dei consumatori, preferenze, ecc...);
- crescita del data warehousing: con la conseguente disponibilità immediata di materiale (databases puliti e ben documentati) su cui effettuare data mining;
- nuove soluzioni IT: le nuove tecnologie, più economiche in termini di archiviazione e di potenza di calcolo, hanno reso possibili progetti di data mining su larga scala, che precedentemente non sarebbero stati realizzabili;
- nuove ricerche sul machine learning: i nuovi algoritmi, studiati nelle università e nei centri di ricerca, trovano immediatamente riscontro nelle applicazioni commerciali.

Il processo di decision making, al giorno d'oggi è molto più complesso che in passato (meno strutturato e con problemi più difficili da affrontare). I decision makers necessitano quindi di strategie e di strumenti in grado di fronteggiare la nuova situazione.

Toward a Definition

E' difficile definire in maniera precisa un'area in continua evoluzione, quale è sicuramente quella del data mining. In assenza di una definizione universalmente accettata, considereremo il data mining come: **“il processo di estrazione di informazione valida, utilizzabile e precedentemente sconosciuta, da grandi databases e l'utilizzo di queste informazioni per prendere cruciali decisioni di business”**. Alcune delle parole utilizzate nella definizione aiutano a chiarire in cosa il data mining si differenzia rispetto ad altre discipline simili, quali il query reporting o l'OLAP:

- informazione valida: un data miner, analizzando larghi insiemi di dati, prima o poi troverà qualcosa di interessante. Contrariamente a quanto dicano i super-ottimisti, è necessario controllare che i risultati ottenuti non siano errati;
- informazione utilizzabile: deve essere possibile tradurre la nuova informazione in un vantaggio di business;
- informazione sconosciuta: il data miner ricerca qualcosa che non è intuitivo ma, anzi, è spesso controintuitivo (più l'informazione si discosta dall'ovvio, infatti, più è grande il suo valore potenziale).

Revolution or Evolution?

Contrariamente a quanto si potrebbe essere portati a credere, il data mining è una disciplina molto più evolutiva, che non rivoluzionaria. I vari filoni che hanno portato su questa strada iniziano negli anni '60 con gli studi di Frank Rosenblatt sul machine learning: l'obiettivo che lo scienziato si poneva era di fare in modo che un computer, partendo dall'osservazione di un certo numero di situazioni conosciute, potesse sviluppare un insieme di regole sottostanti, universalmente vere. Egli sviluppò "Perceptron", predecessore delle attuali reti neurali, su cui si riversarono grandi aspettative, ma che si rivelò un fallimento. Le critiche ed i miglioramenti che ne seguirono confluirono nel 1969 nelle ricerche sulla knowledge discovery. La knowledge discovery (o knowledge engineering) ribalta l'approccio all'apprendimento, in quanto precodifica le regole (o conoscenze) del mondo e le mette successivamente a disposizione della macchina. Nacquero così le prime expert system applications, che vennero inizialmente rivolte ai settori della diagnosi medica e della configurazione dei computer. La gestione della base formalizzata delle conoscenze umane, però, comportava diversi problemi, tra cui gli ingentissimi costi. Inoltre, la limitazione delle tecnologie dell'epoca non consentirono ai sistemi esperti di poter seriamente emulare i ragionamenti/comportamenti di un essere umano. Negli anni '80 ripresero gli studi sul machine learning, con la creazione di nuove e più complesse reti neurali, applicate in nuovi campi come quello della valutazione dei risultati delle campagne di marketing. Contemporaneamente, l'ampia disponibilità di databases commerciali portò alla nascita del database marketing, che consentiva campagne di marketing personalizzate, più mirate alle vere esigenze dei consumatori. Per ultimo, dalla fine degli anni '80 in poi, venne coniato il termine di "knowledge discovery in database", ad indicare il processo generale di estrazione di conoscenza dai databases. KDD è oggi diventato sinonimo di data mining.

What's so Different?

Non è facile vedere la differenza tra il data mining e le altre tecniche di analisi dei dati. In generale, possiamo dire che quando conosciamo chiaramente la forma ed i contenuti approssimativi di ciò che stiamo cercando, non abbiamo a che fare con un problema di data mining. Il data mining elimina infatti i confini di ciò che è possibile scoprire. Esso permette di affrontare domande che escono dal range delle tecniche tradizionali e, ancora più importante, che hanno un valore di business decisamente maggiore.

Tra tutte le varie tecniche di analisi dei dati, è sicuramente la statistica quella che più si avvicina al data mining. E' senz'altro vero anche che la maggior parte di ciò che è possibile fare con quest'ultimo può essere fatto con analoghe tecniche statistiche. Ciò che attrae del data mining, però, è la relativa facilità con cui è possibile ottenere nuove intuizioni (semplicità che non riguarda, però, la successiva fase di interpretazione). Inoltre, questa attrattiva è spiegata anche dai tipi di inputs accettati: non solo dati numerici su cui è necessario applicare forti assunzioni relative alla loro distribuzione (pratica comune nella statistica), ma più vasti insiemi di dati, privi o comunque con poche assunzioni a riguardo. La strategia ottimale rimane tuttavia quella di utilizzare la statistica ed il data mining come due approcci complementari.

Not So Different

Chiarito tutto ciò, il data mining non è altro che l'ultimo di una lunga serie di approcci per la risoluzione di problemi di business mediante l'analisi dei dati. Esso permette miglioramenti significativi nella velocità e nella qualità delle decisioni di business. Come rovescio della medaglia, il data mining porta con sé i problemi comuni a tutti gli approcci basati sui dati, tra cui il "GIGO" (Garbage In Garbage Out): se l'input è spazzatura, anche l'output sarà spazzatura.

The Data Warehouse Connection

Il data warehousing, seppur distinto dal data mining, è strettamente legato ad esso. E' comunque bene ricordare che la presenza di una data warehouse non è un prerequisito fondamentale per poter utilizzare soluzioni di data mining, che vengono spesso applicate a files "lisci", estratti direttamente dalle fonti di dati operazionali.

The Data Warehouse

Scopo di una data warehouse è aiutare a migliorare l'efficacia nel precedere decisioni di business guidate dai dati. Il concetto si basa fundamentalmente sulla distinzione tra dati operazionali (utilizzati per il quotidiano funzionamento dell'organizzazione) e dati informativi (usati per la gestione dell'organizzazione). La data warehouse è progettata per essere un'area di stoccaggio, neutrale, dei dati informativi dell'azienda e per essere utilizzata quale unica fonte di dati di qualità per il decision making.

Una definizione largamente accettata di data warehouse è la seguente: **“una raccolta di dati orientati al business, integrati, variabili nel tempo e non volatili, a supporto delle decisioni del management”**. Anche in questo caso è utile analizzare le parole chiave della dicitura:

- dati orientati al business: presuppone cioè un'attività di data modeling, che traduca i dati in termini di business;
- dati integrati: la data warehouse deve disporre anche dei principali dati provenienti dalle fonti esterne e da quelle operazionali;
- dati variabili nel tempo: i dati contenuti nella data warehouse sono relativi al momento in cui questi vengono inseriti. Essa deve però tenere traccia di tutte le variazioni nei dati, allo scopo di rendere possibili analisi storiche e dei trends;
- dati non volatili: una volta che il dato viene inserito nella data warehouse, questo non deve più essere modificato.

Una ricerca ha stabilito che il 90% delle maggiori organizzazioni sono già in possesso di una data warehouse, o comunque la stanno costruendo.

The Data Mart

Chiaramente, l'implementare una data warehouse non è un'attività banale, specialmente se questa deve essere utilizzata dall'intera impresa. E' per questo motivo che diverse organizzazioni, negli ultimi anni, hanno optato per le data marts, più specialistiche, accessibili e piccole rispetto ad una data warehouse aziendale. Talvolta, le data marts vengono utilizzate anche da quelle imprese che già hanno a disposizione un'ampia data warehouse, allo scopo di favorire processi specialistici.

From Data Warehouse to Data Mine

Le organizzazioni dotate di una più ampia concezione dei sistemi di supporto alle decisioni, stanno ormai passando dal data warehouse al data mine. Si tratta di un passaggio naturale, in quanto queste imprese conoscono il valore strategico del data asset e sono quindi ben disposte nei confronti del data mining. Inoltre, il grosso del lavoro per interpretare e ripulire i dati di business è già stato fatto: il passaggio al data mining consente dunque di capitalizzare ulteriormente gli investimenti riversati

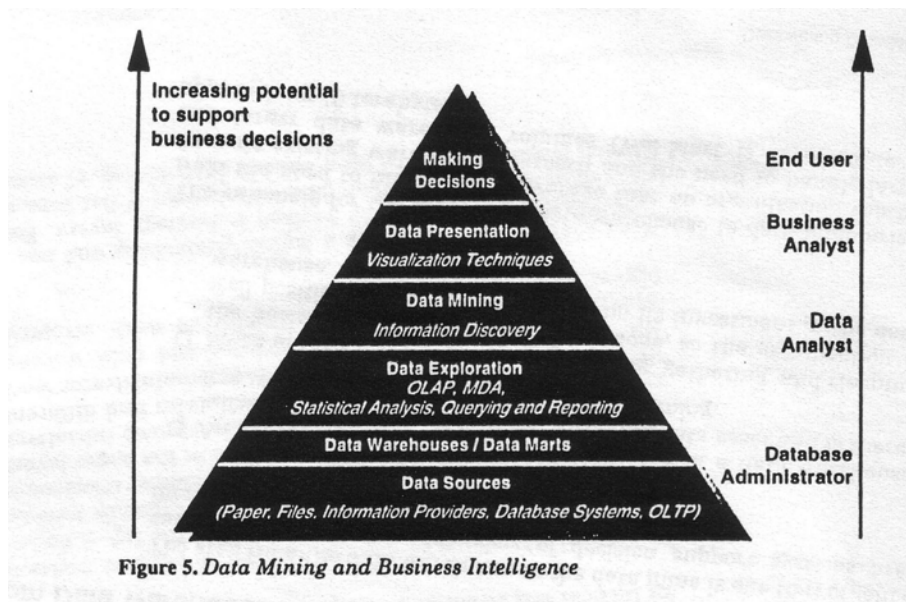
sul data warehousing e di gestire in modo più maneggevole volumi sempre più ampi di informazioni.

From Data Mine to Data Warehouse

Dopo aver implementato una soluzione di data mining, un'organizzazione può scegliere di integrare tale soluzione in un più ampio approccio al decision making basato sui dati, con la creazione di una data warehouse integrata.

Data Mining and Business Intelligence

Utilizzeremo il termine di "business intelligence" come riassuntivo di tutti i processi, le tecniche e gli strumenti che supportano il decision making basato sull'information technology. L'approccio può spaziare dai semplici fogli di calcolo ai più avanzati strumenti di competitive intelligence: il data mining si inserisce in tale contesto come una nuova ed importante componente della business intelligence.



In generale, il valore dell'informazione a supporto del decision making aumenta man mano che si sale dalla base verso la punta di questa piramide. Una decisione basata sui dati, negli strati bassi (dove sono tipicamente presenti milioni di records), influenza generalmente una singola transazione. Al contrario, una decisione basata sui dati degli strati superiori è molto più probabilmente un'iniziativa del dipartimento o dell'intera compagnia. E' diretta conseguenza di ciò il fatto che, ai diversi strati della piramide, troviamo tipologie di utenti profondamente diverse: l'amministratore del database lavora principalmente ai primi due livelli, mentre gli esecutivi dell'azienda si focalizzano sui livelli più alti.

Chapter 2: “Down to Business”

In questo secondo capitolo si porrà l’attenzione sulle applicazioni commerciali di più ampio successo del data mining odierno.

Market Management	Risk Management	Fraud Management
<ul style="list-style-type: none">○ target marketing;○ customer relationship management;○ market basket analysis;○ cross selling;○ market segmentation.	<ul style="list-style-type: none">○ forecasting;○ customer retention;○ improved underwriting;○ qualità control;○ competitive analysis.	<ul style="list-style-type: none">○ fraud detection.

Market Management Applications

L’area applicativa del data mining ormai più diffusa è quella del market management, la cui applicazione più conosciuta è senz’altro quella del database marketing. Obiettivo del database marketing è guidare il marketing e le campagne promozionali mediante l’analisi dei dati contenuti nei databases aziendali, incrociati con informazioni di pubblico dominio. Gli algoritmi di data mining passano al setaccio i dati disponibili, cercando di individuare clusters di consumatori che condividono le stesse caratteristiche (ad esempio gli interessi, il reddito, ecc...). Ogni cluster così individuato può diventare il bersaglio di un sforzo di marketing mirato.

Un’altra area applicativa del data mining è quella della determinazione dei modelli d’acquisto dei consumatori nel tempo. Gli esperti di marketing possono scoprire molte cose osservando semplici operazioni (ad esempio, il passaggio da un conto corrente personale ad uno condiviso indica con buone probabilità il matrimonio di due persone) e predisporre promozioni “just-in-time” ai propri clienti. Ciò consente di migliorare notevolmente il lifetime value di un consumatore.

Terza area applicativa è quella delle campagne di cross-selling, che mirano a rendere un prodotto o un servizio più attraente (ad esempio offrendo buoni sconto o condizioni particolarmente vantaggiose) per i consumatori che acquistano un prodotto o un servizio ad esso associato.

Risk Management Applications

La pratica del risk management non copre solo quei rischi associati ad assicurazioni ed investimenti, ma anche quelli legati al business (minacce competitive, attriti con i clienti, ecc...).

Il data mining è particolarmente indicato per predire le possibili perdite che può provocare il sottoscrittore di una polizza; in questo caso, gli algoritmi permettono inoltre di praticare in tempo reale la tecnica del variabile pricing (costo della polizza differente in funzione della “pericolosità” dell’individuo). Il data mining può anche essere utilizzato per individuare i clienti più propensi a passare alla concorrenza, creando un modello dei consumatori più vulnerabili. I dettaglianti utilizzano le stesse tecniche, mirate però a scoprire la vulnerabilità di certi prodotti alle offerte della concorrenza o alla variazione dei modelli d’acquisto dei consumatori.

Nel settore bancario, una classica applicazione è relativa all’area della concessione dei prestiti: i clienti di lunga data, con un tracciato pluriennale di buoni salari, sono ovviamente preferibili. Relativamente ai prestiti già concessi, il data mining permette invece di segmentare i beneficiari in termini di “tassi di fallimento”.

Oggi più che mai, è vitale per le imprese monitorare la direzione del mercato e ciò che fanno i concorrenti. Questa aumentata attenzione alla competizione ha portato ad un’impennata del ricorso alla pratica della competitive intelligence, il processo di raccolta, analisi e divulgazione delle

informazioni relative agli sviluppi dell'industria o ai trends di mercato, ai fini del miglioramento della competitività aziendale.

Forecasting Financial Futures

La finanza è una disciplina da sempre fortemente interessata al tentativo di predire il futuro. Se le modifiche dei comportamenti finanziari potessero essere predette, infatti, le imprese potrebbero aggiustare le loro strategie di investimento e capitalizzare tali modifiche. L'applicazione di tecniche quantitative, quali la statistica o il data mining, all'area del risk management è spesso indicata dal termine "financial engineering". Una tipica applicazione si ha nel campo dei "futures" (contratti che permettono al possessore di acquistare un certo bene, ad un certo prezzo, in una certa data), dove vengono creati modelli di previsione per il prezzo futuro della commodity in questione, garantendo al broker un notevole vantaggio competitivo.

Fraud Management Applications

La natura umana ci insegna che la frode è inevitabile in qualsiasi tipo di industria, nonostante vi siano alcuni settori dove la probabilità che essa si verifichi è molto più alta rispetto ad altri. Spesso viene usato il data mining: a partire da dati storici viene creato un modello dei consumatori con comportamenti fraudolenti o di quei comportamenti potenzialmente fraudolenti. Il modello così creato aiuta ad identificare istanze di possibili truffe in atto.

Emerging and Future Application Areas

L'attuale interesse nei confronti del data mining commerciale porterà indubbiamente ad un aumento del numero delle aree applicative in cui esso verrà utilizzato. Due delle aree attualmente in sviluppo sono quelle del text mining e del web analytics:

- **text mining:** applicazione delle tradizionali tecniche di data mining ai contenuti, non strutturati, dei databases testuali;
- **web analytics:** utilizzo del data mining in congiunzione ad Internet. L'idea è quella di applicare il data mining ai logs delle attività memorizzati dai web servers, in modo da poter sviluppare conoscenze aggiuntive sul comportamento degli utenti su Internet.

When Things Go Wrong!

Il data mining, se fatto senza riguardo agli obiettivi di business, alla correttezza dei dati ed al senso comune, difficilmente può condurre a risultati soddisfacenti. La pur recente letteratura in materia ci presenta svariati casi in cui queste semplici linee guide non sono state rispettate, con effetti disastrosi sugli indicatori delle performances aziendali.

Chapter 3: “The Data Mining Process”

Before You Start

Siccome non esiste qualcosa che è possibile considerare come un “tipico” progetto di data mining, il processo qui descritto è, per definizione, “generico”.

The Process in Overview

Generalmente, quando la gente parla di data mining si focalizza principalmente sugli aspetti di “mining” e di scoperta. Il mining dei dati, tuttavia, è solo una delle diversi fasi in cui si articola il processo globale (iterativo e “multi-step”) di data mining. A guidare l’intero processo sono gli obiettivi di business: essi costituiscono la base su cui viene costruito il nuovo progetto, i parametri con cui vengono valutati i risultati finali e devono essere un costante riferimento per il team durante le varie fasi di sviluppo.

Nonostante il processo sia altamente iterativo e con molti possibili loopbacks su una o più fasi, possiamo individuare 5 steps sequenziali:

1. **Business Objectives Determination:** definizione chiara del problema di business o della sfida che l’azienda si pone;
2. **Data Preparation:**
 - 2.1. *Data Selection:* identificazione di tutte le fonti di informazione (interne o esterne) e selezione di quel sottoinsieme di dati necessario per l’applicazione di data mining;
 - 2.2. *Data Preprocessing:* studio sulla qualità dei dati, indirizza la futura analisi determinando il tipo di operazioni di mining che è possibile effettuare;
 - 2.3: *Data Transformation:* trasformazione dei dati in un modello analitico. I dati vengono modellati in modo da essere conformi ai formati richiesti dagli algoritmi di data mining e poter così effettuare le analisi precedentemente scelte;
3. **Data Mining:** mining dei dati trasformati nel passaggio 2.3. E’ il cuore del processo, ma, a parte la scelta della combinazione di algoritmi più appropriata, viene svolto in modo completamente automatico;
4. **Analysis of Results:** interpretazione e valutazione dell’output dello step 3. L’approccio all’analisi può variare in funzione dell’operazione di data mining effettuata, ma chiama generalmente in causa qualche tecnica di visualizzazione;
5. **Assimilation of Knowledge:** incorporazione, all’interno dell’azienda e del suo sistema informativo, delle conoscenze acquisite nello step 4.

Non tutte le fasi del processo hanno uguale peso in termini di tempo e di sforzi necessari per il loro compimento, come dimostra il grafico nella pagina seguente.

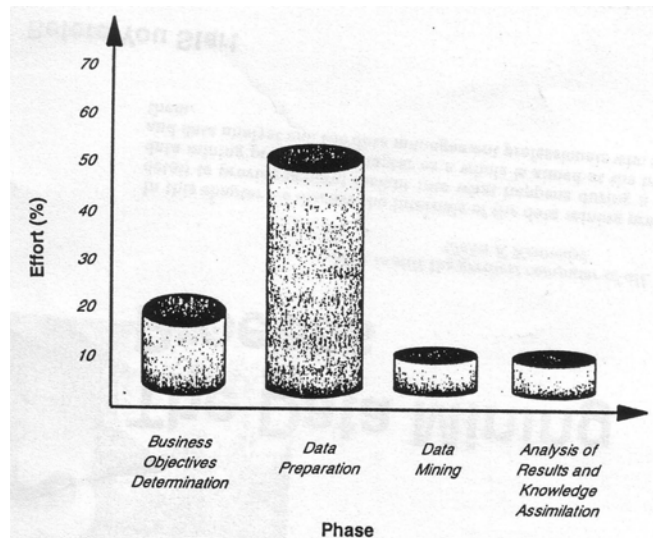


Figure 12. Effort Required for Each Data Mining Process Step

Per la nostra discussione è bene definire i ruoli dei partecipanti ad un progetto di data mining:

- **Business analyst:** dotato di una profonda conoscenza del dominio dell'applicazione, interpreta gli obiettivi e li traduce in esigenze di business, che serviranno per definire i dati e gli algoritmi di mining da utilizzare;
- **Data analyst:** dotato di una profonda conoscenza delle tecniche di analisi dei dati, solitamente ha un forte background statistico. E' in possesso dell'esperienza e dell'abilità necessarie per trasformare le esigenze di business in operazioni di data mining e per scegliere le tecniche di data mining più adatte ad ogni operazione;
- **Data management specialist:** dotato di competenze sulle tecniche di gestione e di raccolta dei dati a partire da sistemi operazionali, databases esterni o datawarehouses.

The Process in Detail

1) Business Objectives Determination

Requisiti minimi di questa fase sono la presenza di un problema o di un'opportunità di business riconosciute come tali ed una "sponsorizzazione" da parte dell'esecutivo dell'azienda. In mancanza di un obiettivo di business da perseguire, molto difficilmente il data mining può portare a risultati concreti. E' in questo step che emergono le aspettative della dirigenza: se troppo elevate possono rapidamente portare al fallimento di un progetto altresì valido. E' essenziale che questo step venga sfruttato anche per una rapida analisi del rapporto costo/possibili benefici.

2) Data Preparation

Si tratta della fase più intensiva del processo, e si articola, come già accennato, in tre stadi (data selection, data preprocessing e data transformation).

2.1) Data Selection

L'obiettivo è quello di identificare le fonti di dati disponibili ed estrarre quei dati necessari per le analisi preliminari in previsione del mining. Ovviamente, la data selection varia in funzione degli obiettivi di business da perseguire. Insieme a ciascuna delle variabili selezionate sono necessarie le

informazioni semantiche corrispondenti (metadati), al fine di meglio comprenderne il significato. Esistono tre tipi di variabili:

- **Categorical:** si tratta di quelle variabili che possono assumere soltanto valori finiti. Esse si dividono in variabili nominali (sulle quali non è possibile effettuare un ranking) e ordinali (in cui, invece, i valori sono ordinabili secondo il loro possibile valore). Ad esempio è una variabile nominale lo status familiare (i possibili valori: “single”, “sposato”, ecc... non sono ordinabili) mentre è ordinale la frequenza d’acquisto (“regolare”, “occasionale”, “rara”);
- **Quantitative:** sono quelle variabili per le quali si è in grado di misurare una differenza tra i possibili valori. Esse si dividono in variabili continue (i cui valori sono numeri reali) e discrete (i cui valori sono, invece, numeri interi). Ad esempio è una variabile continua la media d’acquisto di un prodotto, mentre è discreta il numero di impiegati di un’impresa;
- **Binarie:** le variabili che possono assumere soltanto due valori (tipicamente 0 ed 1).

Dal punto di vista dell’analisi, abbiamo invece:

- **Variabili attive:** quelle scelte per il data mining. Sono chiamate attive perché sono attivamente usate per distinguere i segmenti, fare predizioni o compiere qualche altro tipo di operazione di data mining;
- **Variabili supplementari:** non utilizzate direttamente nell’analisi di data mining, ma utili per la visualizzazione e l’esplicazione dei risultati.

Nel momento in cui vengono selezionati dei dati, un’altra importante considerazione da fare è quella sulla loro “expected shelf life”, ossia il tempo dopo il quale questi dati perdono di valore e devono pertanto essere rimisurati.

2.2) Data Preprocessing

Scopo del data preprocessing è quello di assicurare la qualità dei dati scelti. Dati puliti e ben comprensibili sono infatti un chiaro prerequisito affinché un progetto di data mining abbia successo. Si tratta della parte più problematica del secondo step, in quanto i dati selezionati sono generalmente raccolti da numerosi sistemi operazionali, inconsistenti e poveramente documentati. Il data preprocessing ha inizio con una recensione generale della struttura dei dati e con qualche misurazione della loro qualità. Gli approcci possono essere vari, ma richiamano generalmente una combinazione di metodi statistici e di tecniche di visualizzazione dati.

Per quanto riguarda le variabili categorical, la comprensione è spesso aiutata dalla distribuzione delle frequenze dei valori; graficamente, si ricorre ad istogrammi, grafici a torta, ecc... Quando si ha invece a che fare con variabili quantitative, vengono utilizzate misure quali media, mediana, valore massimo, ecc... Da un punto di vista grafico, invece, tipici strumenti sono i diagrammi di dispersione (scatterplots), che permettono di predire l’andamento di una variabile in funzione del valore di una seconda variabile, e i boxplots, utili per confrontare la media e/o la deviazione standard di due o più variabili.

Durante la fase di preprocessing, due dei problemi che possono più comunemente insorgere sono quelli dei “noisy data” e dei “missing values”:

- **Noisy Data:** una o più variabili hanno valori decisamente diversi da quelli che sarebbe lecito attendersi. Le osservazioni in cui occorrono questi noisy values vengono chiamate “outliers” e la loro presenza può essere dovuta a diversi fattori:
 - errore umano (ad esempio, età di 150 anni invece che di 15);
 - modifiche del sistema operativo che non sono ancora state replicate all’interno dell’ambiente di data mining (ad esempio, la variazione dei codici-prodotto);
 - campionamento inefficace (ad esempio, quando vengono confrontati i redditi di un centinaio di pensionati, ma sono presenti per sbaglio anche un paio di professionisti ancora in attività);

- **Missing Values:** includono valori non presenti o non validi. La mancanza di alcuni valori può derivare da un errore umano, dal fatto che l’informazione non era disponibile al momento dell’input o perché i dati sono stati presi da fonti eterogenee. I data analysts utilizzano diverse tecniche per trattare i missing values:
 - eliminazione delle osservazioni con valori mancanti (problematica nel caso in cui il volume di dati sia basso, oppure quando l’applicazione di data mining è volta alla fraud detection);
 - eliminazione della variabile nel caso in cui, per un significativo numero di osservazioni, manchi proprio il valore di quella variabile (problematica, in quanto la variabile che viene eliminata potrebbe essere molto importante per l’analisi);
 - sostituzione del valore mancante. Per le variabili quantitative, il sostituto può essere il valore più probabile (media o mediana); per quelle categorical, la moda o una nuova variabile (ad esempio: “UNKNOWN”). Un approccio più sofisticato può essere implementato mediante l’utilizzo di un modello predittivo.

Nonostante questo arsenale a disposizione per combattere il problema dei missing values, è bene sottolineare che più congetture devono essere fatte, più ne risentono l’accuratezza e la validità dei risultati del progetto di data mining.

2.3) Data Transformation

L’obiettivo è quello di trasformare i dati preprocessati al fine di generare il modello analitico dei dati. Tale modello è una ristrutturazione consolidata, integrata e time-dependent dei dati selezionati e preprocessati, attinti dalle diverse fonti (operazionali ed esterne). Una volta che il modello è stato costruito, i dati vengono revisionati per assicurarsi che corrispondano ai requisiti degli algoritmi di data mining che devono essere usati.

Le tecniche di data transformation più utilizzate sono:

- **householding:** accorpamento dei registri relativi ai propri consumatori, al fine di costruire un profilo delle proprie relazioni con uno specifico nucleo familiare (tecnica molto utilizzata nell’ambito del customer relationship management);
- **data reduction:** riduzione del numero totale di variabili per il processing, ottenuto mediante la combinazione di più variabili esistenti in una nuova variabile (al di là della difficoltà nell’individuare le variabili che si prestano ad essere combinate, il risultato finale può essere molto difficile da interpretare);
- **data discretization:** conversione delle variabili quantitative in variabili categorical;
- **one-of-N:** conversione di una variabile categorical in una rappresentazione numerica (utilizzata tipicamente nel campo delle reti neurali, per preparare gli inputs).

3) Data Mining

L'obiettivo di questo terzo step è chiaramente quello di applicare gli algoritmi di data mining selezionati ai dati preprocessati. Nonostante, in questo generico processo, la fase di data mining è rappresentata come indipendente, nella realtà essa è praticamente inscindibile dal quarto step (analisi dei risultati), così come è molto raro che essa possa essere ultimata senza tornare, almeno una volta, alla fase precedente (preparazione dei dati).

Ciò che accade durante questa fase varia notevolmente in base al tipo di applicazione che si sta sviluppando: nel caso della segmentazione di un database possono essere più che sufficienti uno o due "passaggi" degli algoritmi sui dati. Situazione ben diversa si ha quando si sviluppa un modello predittivo: il training può richiedere infatti decine e decine di "passaggi".

4) Analysis of Results

E' inutile sottolineare come l'analisi dei risultati del mining sia uno degli steps più importanti dell'intero processo. Il suo obiettivo è quello di rispondere alla domanda: "abbiamo trovato qualcosa di interessante, valido ed utilizzabile?". Mentre le tecniche statistiche si limiterebbero ad un secco "sì/no", i risultati del data mining sono in grado di suggerire la risposta o, nella peggiore delle ipotesi, indicare la direzione da intraprendere in una successiva ricerca.

Nel momento in cui viene sviluppato un modello predittivo, uno degli obiettivi cruciali è quello di testare la sua accuratezza. Molti strumenti di data mining forniscono un grosso aiuto in questo senso, con le "confusion matrixes" (che indicano quanto sono giuste le predizioni sulla base di risultati già noti) e l'"input sensitivity analysis" (che misura l'importanza relativa attribuita a ciascuna variabile in input).

Una delle più comuni fonti di errore, nella costruzione di un modello predittivo, è la scelta di variabili troppo predittive. Un'altra difficoltà è data dall'overtraining: il modello predice bene sui dati utilizzati per il training, ma male su quelli reali. Da tenere in considerazione vi sono poi le cosiddette "association rules": se il livello di confidenza è troppo basso, il modello predittivo individua regole che regole non sono. Viceversa, se il livello è troppo alto, vengono individuate soltanto le regole più generali, già conosciute dagli addetti ai lavori.

Assimilation of Knowledge

Questo step chiude il ciclo con lo scopo di trasformare in azione le nuove informazioni individuate. Le sfide principali da affrontare in questo contesto sono due: presentare le nuove scoperte in maniera convincente e business-oriented; elaborare i modi in cui le nuove informazioni possono essere sfruttate al meglio. La specifica azione di business da intraprendere, ovviamente, varia anche in questo caso in funzione del tipo di applicazione che si è sviluppata e dalle esigenze dell'esecutivo aziendale emerse nel primo step.

Chapter 4: “Face to Face with the Algorithms”

Questo capitolo propone una panoramica delle più comuni operazioni e tecniche di data mining.

From Application to Algorithm

Un novizio può essere confuso dalla moltitudine di termini che circondano il mondo del data mining. Siccome non esiste una terminologia universalmente riconosciuta, utilizzeremo come riferimento questa tabella:

	Market Management		Risk Management		Fraud Management
<i>Business Applications</i>	<ul style="list-style-type: none"> ○ target marketing; ○ customer relationship management; ○ market basket analysis; ○ cross selling; ○ market segmentation. 		<ul style="list-style-type: none"> ○ forecasting; ○ customer retention; ○ improved underwriting; ○ qualità control; ○ competitive analysis. 		<ul style="list-style-type: none"> ○ fraud detection.
<i>Data Mining Operations</i>	Predictive Modeling	Database segmentation	Link Analysis		Deviation Detection
<i>Data Mining Techniques</i>	<ul style="list-style-type: none"> ○ classification; ○ value prediction. 	<ul style="list-style-type: none"> ○ demographic clustering; ○ neural clustering. 	<ul style="list-style-type: none"> ○ associations discovery; ○ sequential pattern discovery; ○ similar time sequence discovery. 		<ul style="list-style-type: none"> ○ visualization; ○ statistics.

Business Applications

Le applicazioni elencate nella figura appena riportata sono quelle tipiche applicazioni di business dove il data mining è utilizzato ai giorni nostri: market management, risk management e fraud management.

Data Mining Operations

Le quattro maggiori operazioni per implementare qualsiasi operazione di business sono:

- **predictive modeling;**
- **database segmentation (clustering);**
- **link analysis;**
- **deviation detection.**

In generale non esiste un legame fisso tra operazioni ed applicazioni (i risultati migliori, talvolta, derivano dall'utilizzo di operazioni non intuitive), anche se, per certe operazioni, questo collegamento è comunemente riscontrato. Le operazioni non sono mutuamente esclusive: nell'ambito del customer retention, ad esempio, l'approccio più utilizzato prevede come prima cosa la segmentazione del database e successivamente l'applicazione della modellazione predittiva ai segmenti più omogenei risultati.

Data Mining Techniques

Le tecniche sono specifiche implementazioni degli algoritmi, utilizzate per portare a termine le operazioni di data mining. La relazione tra tecniche ed operazioni è più forte di quella che intercorre tra operazioni ed applicazioni, ma anch'essa è da ritenersi soltanto una linea-guida.

Non tutti gli algoritmi utilizzabili per implementare una certa operazione sono uguali. Da un algoritmo all'altro possono infatti variare ad esempio il range di input accettati, la trasparenza dell'output del mining, la capacità di gestire ampi volumi di dati, ecc...

Si tratta comunque di un'area in rapida e forte espansione. Al momento, data un certa applicazione o operazione da implementare, non esiste una tecnica migliore rispetto alle altre.

Data Mining Operations

In questa sezione verranno discusse le operazioni di data mining.

Predictive Modeling

Il predictive modeling è simile all'esperienza dell'apprendimento umano, dove usiamo le osservazioni per creare un modello delle caratteristiche essenziali, sottostanti ad un certo fenomeno. Nel data mining, i modelli predittivi vengono utilizzati per analizzare un database esistente, al fine di determinare qualche caratteristica essenziale relativa ai dati. Il modello deve essere in grado di fornire la risposta corretta di fronte ad alcuni casi precedentemente risolti, prima che possa essere utilizzato su nuove osservazioni (approccio "supervised learning"). I modelli predittivi vengo infatti sviluppati in due fasi:

- training: ovvero la costruzione di un nuovo modello sulla base di dati storici;
- testing: cioè il provare il modello su dati nuovi, precedentemente non noti, per determinarne l'accuratezza e le performances del modello stesso.

L'approccio della modellazione predittiva trova la sua più ampia applicazione nei settori del customer retention management, del credit approval, del cross selling e del target marketing.

Esistono due specializzazioni del predictive modeling, che tuttavia condividono lo stesso obiettivo di base (effettuare congetture su variabili di un qualche interesse):

- **classification** (o **categorization**): serve per stabilire una classe di appartenenza specifica per ogni record contenuto nel database, dato un insieme finito e predeterminato di classi;
- **value prediction**: utilizzato per stimare un valore numerico continuo, associato al record di un database (ad esempio, individuare il lifetime value di un nuovo consumatore). Un ulteriore specializzazione del value prediction è lo "scoring", dove la variabile che deve essere predetta è una probabilità od una propensione.

Database Segmentation:

L'operazione di database segmentation (detta anche "clustering") ha lo scopo di partizionare un database in segmenti di records simili (che condividono tra loro un numero di proprietà tale da poterli considerare omogenei), senza nessun intervento da parte dell'utente in merito ai tipi o al numero di segmenti che ci si aspetta di individuare all'interno del database (approccio "unsupervised learning"). Le principali tecniche di database segmentation sono:

- **demographic clustering:** opera principalmente su records con variabili categorical, sfruttando una tecnica di misurazione delle distanze basata sul principio del voto di Condorset;
- **neural clustering:** opera principalmente su input numerici (ma è possibile trasformare gli input categorical in variabili quantitative), sfruttando una tecnica di misurazione basata sulle distanze euclidee.

La segmentazione del database si differenzia dalle altre operazioni di data mining poiché si pone obiettivi molto meno precisi rispetto, ad esempio, al predictive modeling. Per questo stesso motivo, però, essa è più sensibile alle caratteristiche ridondanti ed irrilevanti dei dati (problema tuttavia aggirabile, forzando l'algoritmo di segmentazione a ignorare certi sottoinsiemi di attributi o assegnando un fattore "peso" a ciascuna variabile).

Link Analysis

L'operazione di link analysis cerca di stabilire connessioni (associations) tra records individuali o insiemi di records di un database. Una classica applicazione di questa operazione è l'associations discovery, che ha l'obiettivo di individuare le relazioni tra prodotti o servizi che i consumatori tendono ad acquistare insieme o seguendo una certa sequenza temporale.

Esistono comunque tre specializzazioni della link analysis:

- **associations discovery:** può essere utilizzata per analizzare i beni acquistati nella stessa transazione, in modo da rivelare eventuali affinità nascoste tra i prodotti (market basket analysis o product affinity analysis);
- **sequential pattern discovery:** ha lo scopo di identificare eventuali associazioni tra le transazioni effettuate nel tempo, che possono rivelare informazioni sulla sequenza secondo cui i consumatori acquistano determinati beni o servizi;
- **similar time sequence discovery:** è volta alla scoperta di collegamenti tra due insiemi di dati, rappresentati come serie temporali, ed è basata sul livello di somiglianza tra gli andamenti che entrambe le serie temporali mettono in evidenza. Può essere utilizzata ad esempio per scoprire se l'andamento delle vendite di un prodotto va a contrastare quelle di un altro prodotto.

Deviation Detection

La deviation detection è finalizzata ad individuare eventuali outliers presenti nei dati. Solo in quest'ultimo periodo stanno nascendo i primi algoritmi che automatizzano tale procedura: fino ad oggi, gli analisti hanno eliminato gli outliers aiutandosi con la statistica (in particolare sfruttando il modello della regressione lineare) e con le tecniche di visualizzazione dati rese possibili dalle più moderne tecnologie informatiche.

Data Mining Techniques

Predictive Modeling: Classification

Verranno discusse ora due specializzazioni della classification: la "tree induction" e la "neural induction". Entrambe sono basate su un approccio supervised learning (ovvero il processo automatico di creazione di un modello di classificazione, a partire da un insieme di records

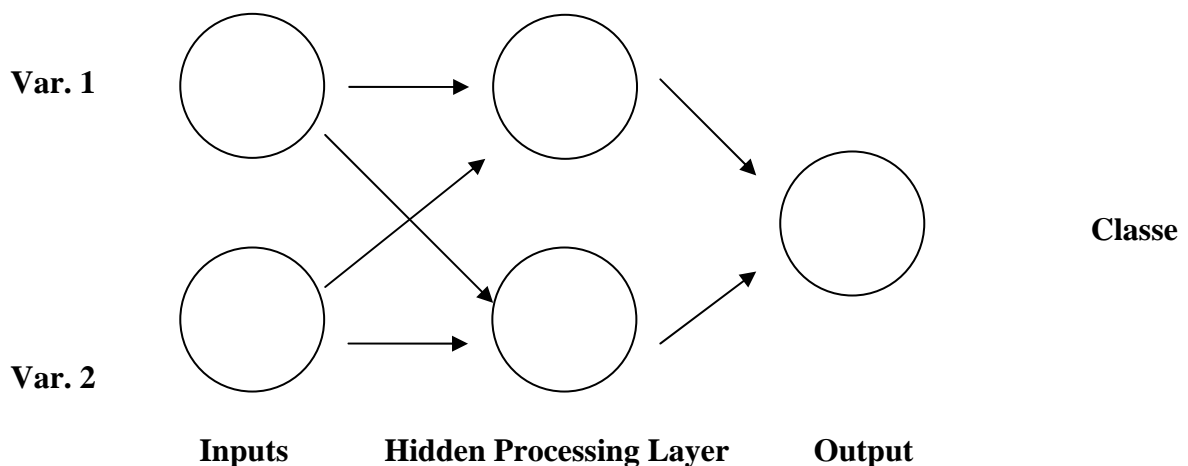
chiamato “training set”). Le tecniche di supervised induction possono essere neurali (rappresentano il modello come un’architettura di nodi e di collegamenti pesati) o simboliche (ad esempio i modelli rappresentati come alberi di decisione o regole IF THEN).

- **Tree Induction:** costruisce un modello predittivo sotto forma di albero di decisione, talvolta di tipo binario (come considereremo nel prosequio). L’algoritmo inizia il suo lavoro identificando la variabile più importante (cioè la più rilevante nel determinare la classificazione) e sulla base di questa divide il database in due parti. L’algoritmo decide quindi sulla seconda variabile, sulla terza, ecc.. fino a quando l’albero non è completamente creato (nel caso in cui le variabili non siano categorical, la divisione avviene tra gruppi di valori). Negli alberi costruiti secondo questi criteri, i punti decisionali vengono chiamati “nodes”, mentre i punti finali, dove sono conservate le osservazioni, prendono il nome di “leaves”.

Tra i vantaggi della tree induction vi sono l’efficienza in termini di tempo di processamento dei dati, l’intuitività e l’utilità insite nella razionalizzazione del processo in esame. Sono però molti i lati negativi, tra cui il fatto che questo metodo si trova in difficoltà nel lavorare con dati continui, può affrontare soltanto quei problemi risolvibili dividendo lo spazio delle soluzioni in aree sempre più piccole e fatica molto nel gestire eventuali valori mancanti;

- **Neural Induction:** rappresenta il modello come un’architettura di nodi e di collegamenti pesati che connettono i nodi e come tale si basa sulle reti neurali. Le reti neurali sono insiemi di nodi, ognuno dei quali connesso con input, output e processing. Tra l’input visibile e lo strato dell’output, vi può essere un qualsiasi numero di strati nascosti di processing. Graficamente, le unità di processing vengono rappresentate con dei cerchi: ciascuna unità di uno strato è connessa con tutte quelle dello strato successivo, evidenziando un valore che esprime la forza della relazione. Questi pesi, rappresentati come linee, sono inizialmente impostati su un valore prossimo allo zero che viene però aggiustato durante la fase di training della rete, in modo che l’output della rete stessa sia conforme alle classi di valori desiderate, calcolate a partire dai dati. Nel caso in cui l’output differisce, viene elaborata una correzione, poi applicata al processing nei nodi della rete. Questi passi vengono ripetuti fino a quando non si raggiunge una delle reimpostate condizioni di stopping (ad esempio, la percentuale dei records classificati correttamente nel corso dell’ultimo passaggio).

Nelle applicazioni commerciali, l’implementazione più utilizzata è quella della “back propagation”. La terminologia fa riferimento alla maniera in cui gli errori vengono propagati (o distribuiti) dallo strato dell’output a quello dell’input durante la fase di sviluppo del modello (applicabile ad un ampio ventaglio di problemi).



L'algoritmo di back propagation si articola in tre fasi:

1. lo strato di input riceve un set di valori numerici: l'input si propaga all'interno della rete fino a raggiungere le unità di output (produzione dell'attuale o predetto modello di output);
2. l'output desiderato è dato come parte dei dati di training. L'output attuale della rete è sottratto dall'output desiderato, producendo un segnale d'errore. Questo segnale è la base della back propagation, che passa gli errori indietro attraverso la rete neurale, forzando un ricalcolo del contributo di ciascuna unità di processing ed estrapolando i corrispondenti aggiustamenti necessari per la produzione dell'output corretto;
3. i pesi delle connessioni vengono aggiustati sulla base di un metodo di minimizzazione dell'errore chiamato "gradient descent" e la rete neurale "impara" così da un'esperienza passata.

L'induzione neurale è più robusta di quella ad albero, in quanto più resistente di fronte ad eventuali dati mancanti. Per contro, essa accetta tipicamente soli input numerici. Entrambe le tipologie di induzione sono comunque soggette al problema dell'overfitting (o overtraining, in sostanza l'ottenere prestazioni migliori con i dati di training, piuttosto che con nuovi valori). Inoltre, per dati impuri o problemi troppo complessi, le reti neurali falliscono nel tentativo di raggiungere un livello di predizione stabile, che corrisponda ai criteri di accettabilità voluti dall'analista.

Molti vedono nell'induzione neurale un approccio black-box alla creazione di un modello. Tale limite viene però aggirato mediante due soluzioni:

- **input sensitivity analysis:** l'analista stabilisce quali sono i campi dell'input più influenti;
- **confusion matrices:** offrono una misura del livello di efficienza del modello di classificazione, mostrando il numero di classificazioni corrette/non corrette per ogni possibile valore della variabile per la quale si sta effettuando l'operazione.

		Predicted		
		Will Leave	Will Stay	
Actual	Left	402	198	600
	Stayed	62	7598	7660
		464	7796	

Le matrici di confusione offrono anche due strumenti per misurare l'efficienza del modello:

- **coverage:** per un dato comportamento, percentuale delle predizioni che il modello ha indovinato, sul totale di tutti gli individui che hanno intrapreso quel comportamento. Ad esempio, se il modello predice che 402 persone abbandoneranno l'impresa e, nella realtà, questo comportamento è seguito da 600 persone, il coverage sarà del 67% (402/600)
- **accuracy:** 100% - per un dato comportamento, la percentuale delle predizioni che il modello ha sbagliato, sul totale delle predizioni fatte per quel

determinato comportamento. Ad esempio, se il modello predice che 62 persone abbandoneranno l'azienda, mentre in realtà non hanno avuto tale comportamento, l'accuracy sarà dell'87% ($100 - 62/464$)

Predictive Modeling: Value Prediction

Le due tecniche tradizionalmente impiegate nel campo del value prediction sono la regressione lineare e quella non lineare. La prima consiste nel tracciare una linea retta nel grafico della distribuzione, in modo tale che la retta sia la miglior rappresentazione della media di tutte le osservazioni in un certo punto del grafico. La regressione lineare, però, funziona bene solo se i dati sono lineari (in caso contrario, l'analista deve aggiustare manualmente l'equazione) ed è molto suscettibile ad eventuali outliers. La regressione non lineare permette di risolvere alcuni dei problemi di quella lineare, ma non è altrettanto flessibile nel gestire tutte le possibili forme del data plot, specialmente nel caso di alti volumi di variabili in input (può risultare pressoché impossibile modellare il tutto con una singola funzione non lineare).

Una nuova tecnica che sta emergendo è però quella delle "radial basis function", più robusta e flessibile rispetto al tradizionale approccio regressivo. L'RBF non funziona scegliendo una singola funzione non lineare, ma la somma ponderata di un insieme di funzioni non lineari (chiamate appunto radial-basis functions). Le funzioni vengono adattate a regioni distinte dello spazio dell'input: per ciascuna di queste regioni viene creata una RBF centrale, che predice la media dei valori nella regione.

Database Segmentation: Demographic Clustering

Il concetto sottostante al demographic clustering è quello di costruire i segmenti, comparando ogni record con tutti i segmenti creati nella fase di data mining. Il singolo record analizzato viene attribuito ad un segmento piuttosto che ad un altro, massimizzando la distanza inter-segmento e minimizzando quella intra-segmento. La tecnica in esame si basa sul principio della votazione di Condorset: i records vengono confrontati, a coppie, su ogni campo: per i campi dove i due records contengono valori uguali viene assegnato un punteggio di +1 (viceversa, -1 per i campi con valori discordanti). La somma complessiva ci dice qual è la somiglianza tra i due records e, di conseguenza, il cluster più appropriato dove inserirli (nel caso in cui il confronto dia esito ad un valore negativo, potrebbe essere creato un nuovo cluster).

Database Segmentation: Neural Clustering

Le tecniche di neural clustering sfruttano le reti neurali basate sulle Kohonen feature maps. Esse consistono in due strati di unità di processing: uno strato di input totalmente collegato ed in competizione con uno di output. Nel momento in cui si inserisce un input all'interno della rete, questo viene confrontato con tutti gli output: quello che risulta più simile all'input viene dichiarato "vincitore" ed il suo collegamento viene pertanto rafforzato. Il peso della connessione viene spostato in direzione del modello in input, di un fattore determinato da un "learning rate parameter" (decrescente all'aumentare degli input forniti alla rete neurale). Vengono inoltre aumentati, seppur in maniera minore, i pesi dei collegamenti delle unità di output adiacenti al vincitore. Si tratta, chiaramente, di un approccio di tipo unsupervised learning.

Link Analysis: Associations Discovery

Lo scopo dell'associations discovery è quello di individuare oggetti che implicano la presenza di altri oggetti nella medesima transazione. Ad esempio, applicare queste tecniche ad un database delle transazioni di un supermercato, può portare a scoprire affinità nell'insieme di articoli venduti.

Queste affinità sono rappresentate da regole di associazione che mostrano, in formato testuale, quali sono quelli articoli che implicano la presenza di altri articoli. Tali regole sono generalmente espresse nel formato “if X then Y”, dove X è detta “rule body” e Y è invece la “rule head” (ad esempio, nella regola: “se un consumatore compra un maglione, allora comprerà anche una camicia”, il maglione è la rule body, mentre la camicia è la rule head). Il problema degli analisti è tuttavia quello di dover dare un giudizio sulla validità e sull’importanza delle regole scoperte. A questo fine vengono utilizzati due parametri:

- **support factor**: indica la ricorrenza relativa della regola di associazione individuata, all’interno dell’insieme complessivo delle transazioni:

$$\frac{nr_transazioni_che_seguono_la_regola}{nr_totale_transazioni} \%$$

- **confidence factor**: indica quanto questa regola è vera nei records individuali:

$$\frac{nr_transazioni_che_seguono_la_regola}{nr_transazioni_che_seguono_solo_la_rule_body} \%$$

- **lift**: indica di quanto aumenta la probabilità di scegliere l’oggetto identificato “rule head”, se l’acquirente ha già scelto quello della “rule body”:

$$\frac{confidence_factor}{support_factor}$$

La tecnica dell’associations discovery è basata sul conteggio di tutte le possibili combinazioni di oggetti. Dal punto di vista della macchina, dunque, vengono utilizzati in sequenza vettori unidimensionali, matrici e modelli a tre e più dimensioni, in funzione del numero di variabili prese in esame. Tale numero influenza ovviamente anche le performance generali degli algoritmi di associations discovery.

Dal punto di vista dell’utente, al contrario, l’associations discovery è molto semplice: è necessario infatti impostare solamente i support ed i confidence factors, ed i risultati guidano ad interpretazioni intuitive. Tra gli svantaggi, vi è l’impossibilità di assegnare valori di business alle associazioni: per gli algoritmi, dunque, non esistono associazioni più importanti di altre.

Link Analysis: Sequential Pattern Discovery

La tecnica del sequential pattern discovery è volta a scoprire modelli tra le transazioni, come la presenza di un insieme di oggetti seguita da un altro insieme di oggetti, all’interno di un database di transazioni temporalmente esteso.

Così come per l’associations discovery, il concetto di **support factor** è importante anche nell’ambito del sequential pattern discovery. In questo contesto, esso indica però la frequenza relativa del modello sequenziale individuato, nell’insieme generale delle transazioni:

$$\frac{nr_consumatori_che_rispetta_la_sequenza}{nr_totale_consumatori} \%$$

La tecnica si basa sul conteggio di ogni combinazione di transazioni che è possibile estrarre dalle sequenze di acquisti dei consumatori e sul mostrare, successivamente, quei modelli sequenziali la cui frequenza relativa è maggiore rispetto al support factor minimo pre-impostato dall’analista.

I vantaggi e gli svantaggi del sequential pattern discovery sono sostanzialmente uguali a quelli dell’associations discovery.

Link Analysis: Similar Time Sequence Discovery

La tecnica del similar time sequence discovery consiste nel cercare tutte le ricorrenze di sequenze simili ad una sequenza data, all'interno di un database di dati memorizzati come serie temporali. Si tratta di una tecnica molto utile per la gestione degli ordini e del magazzino: riscontrando una similitudine nell'andamento degli acquisti di certi prodotti nel tempo, ad esempio, è possibile pianificare in maniera meno dispendiosa gli ordini da inoltrare ai fornitori abituali.

Gli algoritmi, per stabilire se due sequenze sono o meno simili tra loro, utilizzano gli approcci del "margin for error" (la massima differenza per cui i dati di due serie possono essere considerati uguali) e del "permissibile mismatch gap" (il numero di unità temporali consecutive per cui vengono ignorati valori non corrispondenti). Tra gli svantaggi di questa tecnica, vi è sicuramente la sua complessità: i molti parametri da fissare possono infatti essere un ostacolo per i non esperti.

Deviation Detection: Visualization

Le tecniche di visualizzazione sono di gran lunga il più potente sistema per l'identificazione di modelli nascosti all'interno dei dati. Non si tratta di qualcosa di sorprendente, se si pensa che l'80% delle informazioni che un essere umano assorbe proviene proprio dagli occhi.

Esistono svariati modi di visualizzare i dati: per le variabili (e comunque per i dati unidimensionali) si usano tipicamente istogrammi, scatterplots, boxplots e grafici a torta; per i dati a bassa dimensionalità (cioè con al massimo 3 variabili), diagrammi e grafici multidimensionali. Si tratta comunque di strumenti non adeguati per rappresentare dati a più alta dimensionalità, che richiedono diverse modalità di rappresentazione e per le quali è pressoché indispensabile l'aiuto delle tecnologie informatiche. Ad ogni modo, le tecniche di visualizzazione dei dati non sono legate esclusivamente alla deviation detection, ma aiutano anche a meglio comprendere i risultati, ad esempio, di una segmentazione o di un ciclo di associations discovery.

Deviation Detection: Statistics

Mentre le tecniche di visualizzazione possono essere utilizzate per individuare le deviations, la statistica è utilizzata per misurare il loro livello di significatività (o interesse). La statistica è altresì utile nella fase di data preparation, per capire la natura dei dati (ivi inclusa l'eventuale presenza di outliers), ed in quella dell'analysis of results, per testare ad esempio la qualità dei segmenti prodotti da una database segmentation (mediante una misurazione dell'omogeneità interna ad ogni singolo segmento).