UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

TESI DI LAUREA TRIENNALE

IN STATISTICA E GESTIONE DELLE IMPRESE

STRUMENTI INNOVATIVI PER L'ORGANIZZAZIONE DEI DATI AZIENDALI

Relatore: Ch.mo Prof. Susi Dulli

Laureando: Mauro Santagiuliana

ANNO ACCADEMICO 2003-2004

1. BUSINESS INTELLIGENCE

1.1 DEFINIZIONE DI BUSINESS INTELLIGENCE

Per Business Intelligence o Intelligenza Gestionale detta anche DSS o sistema di supporto alle decisioni si intende un processo di trasformazione dei dati grezzi in informazioni al fine di ottenere un vantaggio di business.
Per ottenere questo processo i dati devono essere organizzati in sistemi strutturati per il supporto decisionale, sistemi detti Data Warehouse.
I manager e le persone interessate al processo di prendere decisioni hanno bisogno di accedere ai dati del Data Warehouse in un modo che questo supporti la loro visione globale dell'azienda. Il maggior pregio della Business Intelligence non viene però dalla semplice raccolta dei dati e dal loro immagazzinamento in Data WareHouse, ma piuttosto deriva dall' uso di strumenti di Business Intelligence (BI) per analizzare ciò che nascondono i dati e ricavarne delle informazioni. Infatti il Data Warehouse (DW) è solo una struttura che permette di riunire in maniera consistente la grossa mole di dati, ma non possiede i mezzi reali per analizzarli ed è solo la combinazione del DW con gli strumenti della BI che permette questo.
Questi strumenti devono quindi essere in grado di fornire in tempo reale informazioni, rapporti e consentire analisi di varia natura (What If Analisys, On Line Analitical Processing, Data Mining).
La What If Analisys permette previsioni basate su ipotesi sui dati futuri: ad esempio possiamo prevedere cosa succede alla vendita di coperchi se applichiamo uno sconto del 5% alle pentole smaltate.
L'OLAP mette a disposizione del manager un ambiente di dati multidimensionale, nel quale può eseguire ricerche aggregando i dati in suo possesso: è possibile ad esempio ottenere informazioni sulle vendite di prodotti alimentari avvenute in veneto nell' ultimo mese coinvolgendo le dimensioni di tempo, luogo e prodotti.
Il Data Mining invece applica tecniche di varia natura (statistica e di intelligenza artificiale) agli archivi aziendali alla ricerca di quelle informazioni che non sono visibili in un primo istante perché immessi in una quantità enorme di dati simili.

2. DATAWAREHOUSE

2.1 DEFINIZIONE DI DATAWAREHOUSE

La definizione del termine Data WareHouse risale ai primi anni 80 e con esso si intende una collezione di dati a supporto delle decisioni **integrata**, **non volatile**, **orientata ai soggetti** e **dipendente dal tempo**.
L' **Integrazione** è l'aspetto più importante di un DW e con essa si intende quel processo secondo cui i dati che nei diversi sistemi operazionali sono memorizzati in maniera non omogenea vengono codificati nel data warehouse in una maniera unica e consistente. Ad esempio si suppongano quattro diversi sistemi operazionali che possono avere quattro differenti fonti dati che identificano il sesso : nell'applicazione A (m,f), nell'applic. B (0,1), nell'applic. C (x,y) e nell'applicazione D (maschio,femmina). Bisogna quindi decidere quale forma vogliamo tenere come valida e codificare questa nel Datawarehouse. Lo stesso vale per le strutture delle chiavi, per la misura degli attributi e per le caratteristiche fisiche dei dati.
Per **non volatile** si intende che i dati contenuti nel DW non possono essere modificati come in un sistema informativo normale. Le operazioni che avvengono in un DW sono infatti quelle di un caricamento iniziale(solitamente in massa), e di un'aggiunta su base periodica di altri dati. Non vengono effettuati aggiornamenti (nel senso generale) e le uniche cancellazioni sono quelle effettuate su dati errati.
Per **orientato ai soggetti**, si intende che i dati vengono organizzati in base ai soggetti principali del business (il prodotto, il cliente, l'agente, l'attività, ecc.) e non in base ai processi e/o alle funzioni aziendali (la fattura, la riga di prima nota, la riga d'ordine, ecc.), come avviene nei sistemi operazionali.
Per **dipendente dal tempo** si intende che a differenza dei sistemi operazionali i dati sono storici. Nel Data Warehouse infatti è possibile tener conto della storia dei soggetti , effettuare degli snapshot (fotografie istantanee) in ogni momento e nella struttura delle chiavi si fa riferimento al tempo. Quando si modifica un dato nei sistemi operazionali, si perde il riferimento al dato precedente ovvero i dati presentano sempre il valore corrente cosa che invece non avviene nel DW. Ad esempio se al tempo t1abbiamo quattro record del tipo:

a	t1	fattura01
b	t1	fattura02
c	t1	fattura03
d	t1	fattura04

e vogliamo modificare al tempo t2 il record con chiave "a" da fattura01 a fattura05 si avrà:

a	t1	fattura01
a	t2	fattura05
b	t1	fattura02
c	t1	fattura03
d	t1	fattura04

2.2 DAI SISTEMI OPERAZIONALI AL DATAWAREHOUSE

Per molto tempo si sono utilizzati sistemi di gestione di database di tipo operazionale o transazionale in cui i dati erano orientati all'applicazione, dettagliati (basso livello di granularità), potevano essere aggiornati e vi era il supporto di operazioni giornaliere necessarie alla gestione dell'azienda. Da qualche tempo si è avuta un'evoluzione della tecnologia che ha portato a strumenti (i Data Warehouse) dove i dati sono orientati ai soggetti, si presentano in una forma "sommarizzata" che va incontro alle necessità della gestione dell'azienda (alto livello di granularità), non possono essere aggiornati (ovvero non perdono il riferimento al dato precedente) e si presentano varianti nel tempo. I dati presenti nel DW vengono forniti dai sistemi operazionali e da altre fonti esterne e vengono caricati attraverso procedure di ETL (Extraction, Transformation and Loading) tramite le quali viene effettuata una sorta di "pulizia".
Quando furono costruiti i sistemi operazionali nessuno pensava ad una loro possibile integrazione futura. Ogni applicazione aveva il suo set personale di requisiti e non vi era considerazione per le altre applicazioni durante lo sviluppo del processo. Gli stessi dati quindi si presentavano in locazioni differenti con nomi diversi, qualcuno era etichettato con lo stesso modo in posti diversi e qualche altro in tutte le locazioni con lo stesso nome ma riferito a differenti misure e cosi' via. Diventa quindi necessario un filtraggio di questi dati disomogenei provenienti dai sistemi operazionali per presentarli in maniera omogenea e comprensibile all' interno del Data Warehouse.
Si deve tenere presente che i modelli di dati dei sistemi operazionali hanno caratteristiche diverse e a volte contrastanti rispetto a quelli dei sistemi decisionali dei Data WareHouse.
Ne è un esempio il concetto di normalizzazione ampiamente utilizzato nei primi per eliminare le ridondanze. Le tabelle che si presentano in questo contesto sono infatti normalizzate e numerose, mentre in un DW si presentano denormalizzate e poche. Qui inoltre vengono distinte in tabelle dei fatti e tabelle delle dimensioni.
La normalizzazione determina una forte frammentazione nello schema dei dati e questa in contesti decisionali potrebbe avere un impatto fortemente negativo in quanto aumentando il numero di tabelle, aumenta la complessità di join (collegamenti) tra queste tabelle e diminuisce la prestazione del sistema. Considerato inoltre che la ridondanza dei dati genera un modello più

semplice, la denormalizzazione sembra adattarsi meglio al contesto decisionale.
Vi sono due particolari modelli a cui si ricorre nel contesto decisionale: lo **Star Schema** e lo **Snowflake Schema**.
Il nome Start Schema deriva dalla somiglianza della sua immagine con una stella: abbiamo infatti una grande tabella dei fatti centrale(es. Vendite) ed un insieme di tabelle satellite (tabelle delle dimensioni) più piccole(es. Prodotti, Clienti, Tempo, Geografia). La tabella dei fatti è l'unica a possedere collegamenti multipli (più chiavi esterne) ed ha un volume molto maggiore (come quantità di informazioni contenute) di quello delle tabelle delle dimensioni.
Le tabelle satellite hanno un solo collegamento (tramite la chiave) con la tabella dei fatti (questo minimizza il numero delle join necessarie per le query) ed ognuna rappresenta una dimensione. Esse sono molto più piccole (in termini di dimensioni) della tabella dei fatti. In genere, solo la tabella dei fatti è normalizzata mentre le tabelle delle dimensioni sono generalmente denormalizzate.
Lo SnowFlake-Schema assomiglia invece ad un fiocco di neve. Qui, rispetto allo Star-Schema, le tabelle delle dimensioni vengono ulteriormente frammentate.

2.3 TECNOLOGIA DEL DATA WAREHOUSE

Un DW è solitamente, ma non necessariamente, una piattaforma hardware le cui dimensioni possono essere ampie (un mainframe) o ridotte (una postazione di lavoro) o in alcuni casi si può trattare di un insieme di piattaforme distribuite o di una serie di nodi nell'ambito di una piattaforma di più ampie dimensioni.
All'interno dei Data WareHouse si può fare una distinzione fra il DW aziendale e il Data Mart.
Il primo viene utilizzato da più reparti all'interno di un'azienda mentre il secondo o DW tematico è più specifico dato che viene utilizzato da un reparto o da un gruppo di utenti aziendali per svolgere un determinato tipo di compiti.
Quest'ultimo inoltre viene suddiviso in Data Mart dipendente e Data Mart indipendente a seconda che contenga o meno informazioni replicate (ma con la stessa codifica) in altri Data Mart. Successivamente vengono presentate nelle figure le tre diverse architetture.

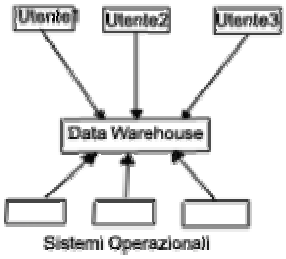


Fig.2.1

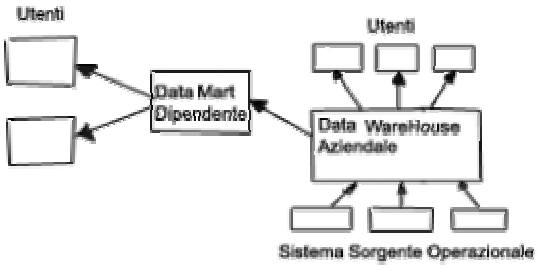


Fig. 2.2

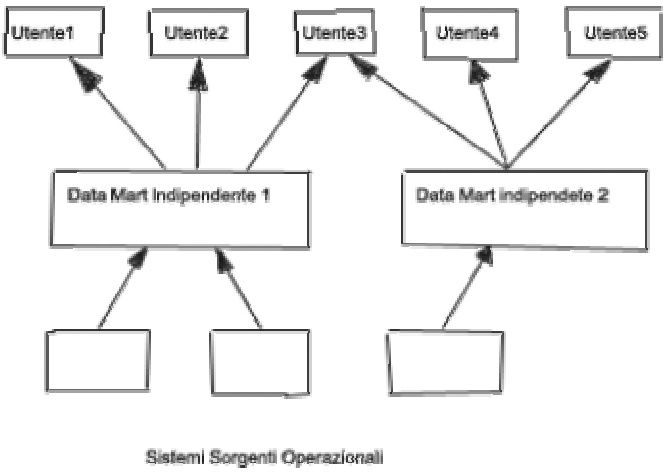


Fig. 2.3

In Fig. 2.1 si rappresenta un DW alimentato da tre distinti sistemi operazionali in cui gli utenti appartengono ai diversi reparti dell'azienda.
In Fig. 2.2 vi è un Data Mart dipendente che utilizza un sottoinsieme di dati del DW aziendale mentre in Fig. 2.3 il Data Mart è un sistema autonomo.

Notare che l'utente 3 accede ad entrambi i Data Mart potendo essere ad esempio un utente interessato a confrontare i dati risultanti dalle elaborazioni effettuate nei due Data Mart.
Un altro aspetto da considerare è che vi possono essere DW locali e DW distribuiti. I primi contengono dati di interesse solo locale e non vi è coordinazione di dati o strutture dati tra un DW locale e un'altro, i secondi contengono dati e strutture di interesse locale (nei vari Dw locali) e comuni a tutta l'organizzazione (nel DW globale) (vi è cioè un Dw globale e più Dw locali).
Dalla definizione di Data Mart nasce la possibilità di procedere, nella costruzione di un DW, in due modalità differenti:
Top-down: si costruisce una base informativa pulita ed integrata a livello aziendale da cui deriveranno diversi Data Mart. Il risultato è un ingente sforzo iniziale ma anche una solida base per l'evoluzione futura.
Bottom-up: si realizzano prima i Data Mart i quali, nel loro insieme, costituiranno il DW. Il risultato è un investimento iniziale ridotto con un ritorno immediato. L'aspetto negativo sta nella complessità e nel costo del consolidamento del Data Mart se la progettazione non ha tenuto conto del contesto più ampio rispetto allo spazio del problema che si stava considerando.

2.4 I METADATI

Il termine metadati indica i "dati relativi ai dati" e con esso si identifica un sistema che documenta i dati di un database permettendo all'utente finale (inteso come analista della DSS) di cogliere il reale significato di nomi, codici postali, indirizzi e altri elementi informativi.
Attraverso i metadati, l'utente può facilmente "navigare" attraverso l'informazione in quanto essi suggeriscono se ciò che si sta cercando esiste e in quale tabella o database si può reperire e come sono stati memorizzati i dati.
Per questo motivo, mentre nell'ambiente operazionale i metadati vengono considerati poco importanti e utilizzati in maniera opzionale, nel Data Warehouse essi acquistano un ruolo centrale in quanto vengono usati all'inizio di ogni analisi dall'analista del DSS e in maniera ricorrente.
Un'altra ragione dell'importanza dei metadati è che essi tengono traccia dei dati sorgenti nel meccanismo di mapping (conversione, filtraggio, riepilogo, cambiamenti strutturali) tra l'ambiente operazionale e l'ambiente del Data Warehouse. Inoltre, poiché il lasso temporale di vita di un Data Warehouse è di 5/10 anni ,la traccia della sua struttura può essere recuperata dopo questo tempo proprio dai metadati.
I metadati possono essere classificati in metadati tecnici e metadati di business.

- Metadati Tecnici
Sono utilizzati da analisti e sviluppatori per realizzare e gestire il DW. Questo tipo di metadati riguarda per esempio i legami tra sistema operazionale e il sistema decisionale, l'origine dei dati, la struttura del DataBase decisionale (nomi, formati, dimensioni ecc.) la frequenza e modalità di aggiornamento e archiviazione dei dati.
- Metadati di business
Sono utilizzati per supportare gli utenti nella corretta interpretazione dei dati di business. Questi metadati definiscono il collegamento tra oggetti informatici e concetti propri dell'attività di business ed è quindi attraverso di loro che si viene a creare un canale di comunicazione tra utenti finali e personale dei sistemi informativi.

2.5 OLAP

L'OLAP (On Line Analytical Processing) è una tecnologia intesa come insieme dei sistemi di analisi dei dati pensati e ottimizzati per garantire la massima performance e la massima "estensione" delle interrogazioni.

Una definizione della funzionalità OLAP è stata proposta da M.Pendle e R.Creeth (1995), con l'acronimo **FASMI** (Fast Analysis of Shared Multidimensional Information). (1)

Fast: indica che obiettivo fondamentale del sistema è la velocità di risposta, dell'ordine di pochi secondi, essendo il sistema stesso concepito per funzionare in modalità interattiva.
Analysis: significa che l'applicazione deve essere in grado di svolgere qualsiasi analisi sia esso di tipo logico, di tipo statistico-operativo o semplicemente di aggregazione e presentazione via report.
Shared: esprime l'esigenza di un uso condiviso dell'applicazione.
Multidimensional: rappresenta l'essenza dei prodotti OLAP, il cui obiettivo primario è fornire una vista multidimensionale dei dati.
Information: pone l' accento sull'esigenza di accesso indiscriminato ai dati, indipendentemente dalla piattaforma fisica su cui essi risiedono e sul formato di codifica.
L'OLAP infatti è uno strumento multiutente e ad alta velocità che opera su strutture dati multidimensionali che sintetizza in cubi logici indicizzati su più assi dimensionali.
Le operazioni più comuni che si effettuano con l'OLAP sono:
Roll-up o drill-up che è un'operazione che aggrega i dati.
Drill-down che disaggrega i dati.
Slice: permette di vedere il cubo dimensionale trasversalmente a fette. Per ottenere tale effetto si fissa un valore numerico all'interno di almeno una dimensione e si analizzano i dati rispetto a tutte le altre.

Dice: permette di vedere il cubo dimensionale attraverso i suoi sotto-cubi.
Per ottenere tale effetto si fissa un intervallo su ciascuna dimensione ottenendo cosi' una riduzione volumetrica senza però diminuire il numero delle dimensioni considerate.
Rotate (Pivot): consiste nella possibilità di ruotare le dimensioni del cubo per vedere il fatto di business secondo ottiche diverse.

I prodotti OLAP si suddividono in varie categorie :ROLAP o relational OLAP, MOLAP o multidimensional OLAP e HOLAP o Hybrid OLAP che è un'integrazione tra MOLAP E ROLAP.
Se i dati vengono gestiti da un Database relazionale si ha un ROLAP (es. Oracle 9i+Discoverer) se invece è supportato da un Database multidimensionale si ha un MOLAP (es. Express Server). L' HOLAP è invece supportato o da un database relazionale o da un database multidimensionale.

3. DATA MINING

Le analisi che si compiono nell'ambito del DSS si suddividono in:

1. livello query standard; (OLTP o ON LINE TRANSACTION PROCESSING)
2. livello analisi multidimensionali; (OLAP)
3. livello analisi statistiche o livello modellazione/segmentazione;
4. livello Knowledge Discovery (scoperta della conoscenza) in Data Base (KDD).

Le query rappresentano il tipo di analisi più diffuso e più semplice da effettuare da parte dell'utente e un esempio può essere quello di visualizzare tutti i clienti che lo scorso anno hanno utilizzato il prodotto y. L' output è in questo caso un report, mentre l'interrogazione è una richiesta di informazione. L' ipotesi che sta alla base di questo metodo è forte, la complessità è bassa ed i tempi di risposta sono brevi.
L'analisi multidimensionale rappresenta il gradino successivo e si tratta di un tipo di analisi più dettagliata della precedente. L' ipotesi che sta alla base è medio bassa ed i tempi di risposta sono generalmente bassi.
Questa analisi viene fatta con pacchetti software ad hoc attraverso cui l'utente interroga gli archivi. Le analisi statistiche permettono di indagare ancora più a fondo i dati. L' ipotesi che sta alla base di questo tipo di analisi è leggera, la complessità è medio alta ed i tempi di risposta lunghi. Per realizzarla servono specifici pacchetti software di tipo statistico e viene effettuata da utenti "esperti" quali analisti di marketing e statistici.
Nella KDD infine algoritmi molto potenti cercano modelli nascosti o particolari ralazioni nei dati che non sono specificati a priori come nella modellazione/segmentazione. La complessità nell'analisi è massima e i tempi di elaborazione molto elevati.

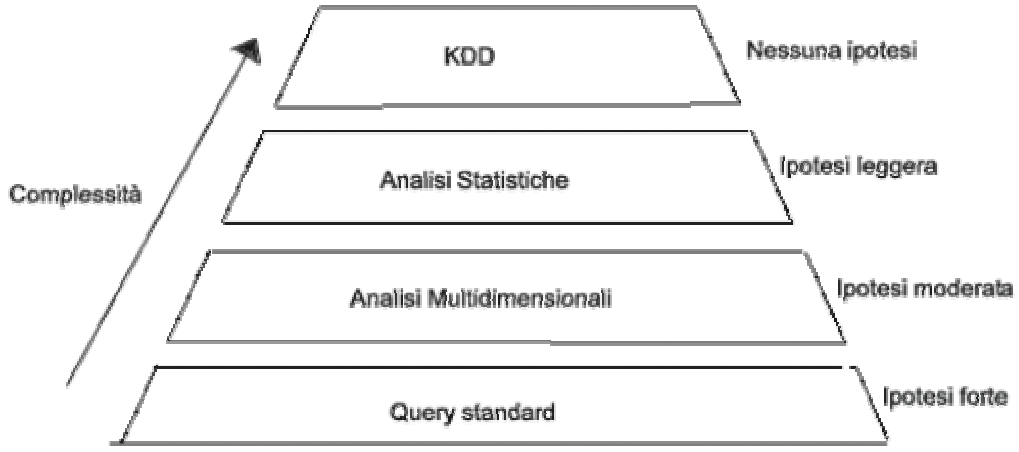


Fig 3.1

3.1. DEFINIZIONE DI DATA MINING

La categoria di analisi che comprende sia il livello 3 che il livello 4 è il Data Mining ovvero quel particolare insieme di tecniche di analisi ripartite in varie fasi procedurali volte ad estrarre conoscenze non note a priori da grandi insiemi di dati apparentemente non correlati.
In altre parole, col nome Data Mining si intende l'applicazione di una o più tecniche che consentono l'esplorazione di grandi quantità di dati, con l'obiettivo di individuare le informazioni più significative e di renderle disponibili e direttamente utilizzabili nell'ambito del processo di decision making (prendere le decisioni).
I processi di analisi dei dati hanno subito una notevole evoluzione negli ultimi 30 anni. Negli anni 60 si utilizzavano sistemi che producevano report standardizzati, che contenevano semplici informazioni riassuntive o prestabilite. Negli anni 80 fu introdotta la possibilità di eseguire interrogazioni differenziate su database rendendo più facile l'identificazione di andamenti relativi, per esempio, a un certo prodotto o a una certa area geografica.
All' inizio degli anni 90 lo sviluppo del software di analisi ha puntato alla possibilità di scavare nei propri dati in tempo reale. Per esempio guardando ad una tabella relativa alle vendite ripartite per zona e prodotto, l'utente può selezionare una zona per vedere l'andamento a livello di singola regione o provincia.

Gli strumenti attuali tendono ad implementare la possibilità di passare al setaccio i dati per scoprire relazioni significative. Sono questi gli strumenti del Data Mining.
La differenza con gli strumenti OLAP è che mentre il Data Mining indica il **perché** un certo fenomeno sta succedendo, l'OLAP si limita a dire **cosa** sta succedendo. Infatti nell'OLAP le relazioni fra i dati e le linee di tendenza sono spesso nascosti nei report riassuntivi ed è proprio il Data Mining che aiuta le aziende a scoprire queste preziose informazioni. E' 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 Data Mining.

3.2 TECNICHE DI DATA MINING

Alcuni strumenti di data mining sono: 1) le tecniche di visualizzazione dei dati che creano grafici multidimensionali al fine di identificare relazioni complesse e individuare l'informazione nascosta, 2) gli alberi decisionali, 3) le reti neurali, 4) la cluster analysis, 5) l'analisi fattoriale e 6) le associazioni e le sequenze.
Gli alberi decisionali sono rappresentazioni grafiche costruite suddividendo ripetutamente i dati secondo sottogruppi definiti dai valori delle variabili di risposta, per trovare sottoinsiemi omogenei. Tale suddivisione produce una gerarchia ad albero, dove i sottoinsiemi vengono chiamati nodi e quelli finali foglie. Vantaggio è l'interpretazione immediata dei risultati,ma svantaggio è la tendenza dell'albero a raggiungere dimensioni notevoli anche se in questo caso si tende a scomporlo in più sottoalberi.
Le reti neurali sono modelli che simulano la struttura del cervello umano imitandone i meccanismi di apprendimento. In base ai dati di input le reti neurali correggono i parametri del modello per trovare relazioni tra i dati.
Le reti permettono di classificare dati, costruire modelli predittivi, segmentare e prevedere gli andamenti futuri.
La cluster analysis è una tecnica di riduzione dei dati che raggruppa casi o variabili in condizioni di similarità.
L'analisi fattoriale è un'altra tecnica di riduzione che ricava fattori detti anche "variabili latenti" che concentrano le informazioni contenute originariamente in un numero elevato di variabili.

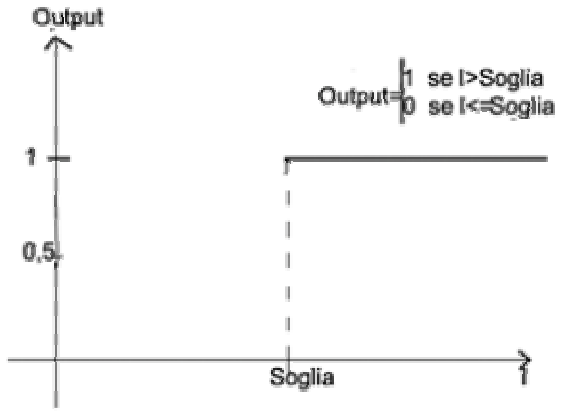
Un altro strumento sono le associazioni e le sequenze. Le regole di associazione spiegano le relazioni tra informazioni che si presentano assieme in un evento. L'applicazione principale di queste regole è la Market Basket Analysis (MBA) spesso applicata in ambito retail. Nella MBA si studiano i panieri d'acquisto con lo scopo di individuare regolarità negli acquisti e relazioni non note o banali nei dati.
Le regole hanno la forma condizione 1-condizione2-cond.ne n-conclusione

Associati ad ogni regola vi sono poi una serie di indicatori come il supporto e la confidenza. Il supporto indica la frequenza assoluta del verificarsi della regola rispetto a tutte le transazioni presenti nel dataset. La confidenza è un rapporto che indica quante volte in percentuale si verifica la conclusione quando sono verificate le condizioni. Un esempio di regola può essere la presente: Gelati-birra⇒Pane 1550(35%,70%)
Questa regola dice che tra tutte le transazioni all'interno del dataset nel 35% dei casi (1500 transazioni) compaiono gelati birra e pane e nel 70% dei casi in cui vengono acquistati gelati e birra è presente anche l'acquisto del pane.
Non c'è un criterio standard per definire "interessante" una regola in quanto si possono fare valutazioni sia sul supporto che sulla confidenza e comunque occorre una buona conoscenza del fenomeno studiato per dividere regole imprevedibili e interessanti da regole note che non portano alcuna informazione aggiuntiva.
Lo studio delle sequenze è simile a quello delle associazioni ma in questo caso è presente una componente temporale e invece che considerare eventi simultanei vengono presi in esame successioni di eventi. Un esempio di questo nell'ambito del retail è:
frutta⇒yogurt⇒latte 1200(20%,35%)
In questa regola si dice che in 1200 acquisti (20% del totale) avviene l'acquisto di frutta seguito dall'acquisto di yogurt e poi dall'acquisto di latte.
Nel 35% dei casi in cui all'acquisto di frutta si è seguito l'acquisto di yogurt, si ha avuto acquisto di latte.
Le tecniche di clustering e l'uso delle reti neurali non supervisionate consentono di effettuare operazioni di segmentazione sui dati, cioe` di individuare gruppi omogenei (o tipologie omogenee), che presentano regolarità al loro interno in grado di caratterizzarli e differenziarli da altri gruppi.
Le reti neurali e gli alberi di decisione consentono di effettuare operazioni di classificazione: fanno cioè uso della conoscenza acquisita in fase di addestramento, durante l'esplorazione dei dati contenuti nel database, per classificare nuovi oggetti o prevedere nuovi eventi.
Un esempio di approccio pratico al data mining è la metodologia S.E.M.M.A. (Sample, Explore, Modify, Model, Assess) proposta da SAS Institute: la nuova informazione viene "scoperta" nelle fasi di esplorazione e di modellistica, mentre le altre fasi sono costituite da attività di rifinitura e supporto. Questa metodologia si configura come un vero e proprio processo dove ogni fase è caratterizzata da un input e da un output, che diventa input per la fase successiva.
Nella fase Sample si estrae una campione di dati dal data set abbastanza grande per contenere ancora informazioni significative ma abbastanza piccolo per analizzarla velocemente. L'esplorazione dei dati serve per cercare relazioni e anomalie nei dati e per capire quali possono essere quelli di interesse. La fase Modify serve per creare, selezionare e trasformare le

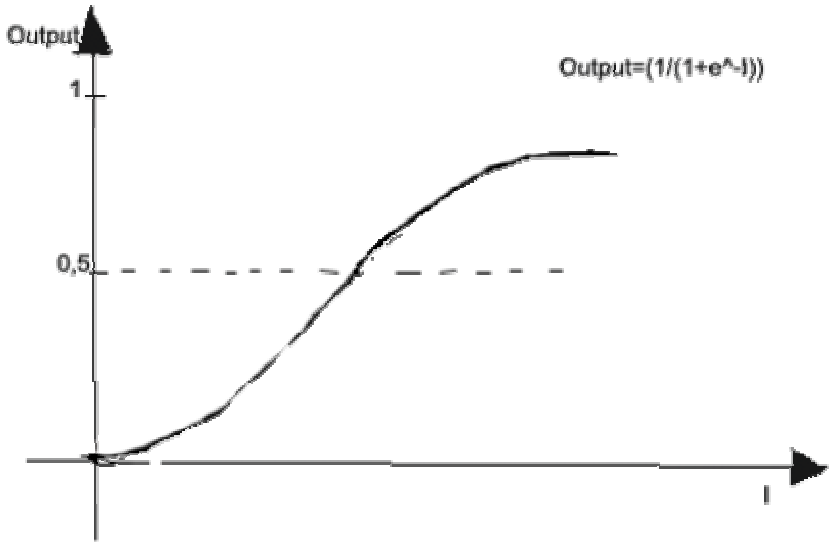
variabili e le misure per mettere a punto il processo di costruzione del modello. Vengono poi ricercate automaticamente le variabili significative ed i modelli che forniscono le informazioni contenute nei dati (Model). Infine, nella fase Assess, si valutano l'utilità e l'affidabilità delle informazioni scoperte tramite la verifica di ipotesi ricavate su altri campioni ed i metodi del test statistico.

3.3 Reti Neurali

Le reti neurali sono tecniche analitiche nate dopo la scoperta dei processi di apprendimento nei sistemi cognitivi e delle funzioni neurologiche del cervello: esse sono in grado di prevedere nuove osservazioni a partire da altre osservazioni dopo aver eseguito un processo cosiddetto di apprendimento a partire da dati preesistenti.
La struttura di una rete neurale è quella di essere un semplice modello del cervello costituito da neuroni e da connessioni tra essi. La funzione di un neurone è quella di sommare i suoi input e produrre un output qualora tale somma sia maggiore di un dato valore di soglia. Questi output vengono poi trasmessi ad altri neuroni tramite giunzioni buone (segnale trasmesso alto) o cattive (segnale trasmesso basso). Questa efficienza delle giunzioni è modellata considerando un fattore moltiplicativo (peso) per ciascun input del neurone (una buona sinapsi avrà quindi un peso maggiore di una cattiva). Un neurone calcola quindi la somma pesata $I = \sum Wi*Xi$ dove Wi è il peso dell'i-esimo neurone mentre Xi è l'input i_esimo. L'output del neurone avviene se la somma I è maggiore di un valore di soglia fissato ad esso ed è dato dalla trasformazione dell' input I tramite una funzione di attivazione. Tale funzione può essere :
-a gradino



- con la curva sigmoide



Il principio di apprendimento permette che la rete impari dai suoi errori. All'inizio i pesi sono scelti a caso, poi le reti li modificano fino ad assegnare loro quei valori che consentono di rispondere in modo desiderato ad una certa stimolazione esterna. Tutti i metodi di apprendimento si dividono in due classi: supervisionato e non-supervisionato.
Il più considerato è il supervisionato in cui la rete si modifica di volta in volta in base a quello che dice un "insegnante" esterno finchè produce un output corretto. Un esempio si appendimento supervisionato è quello di Back propagation (propagazione dell'errore all'indietro): la rete calcola per ogni output la differenza tra stato di attivazione prodotto dalla rete e quello stabilito dall'input. Questo serve a modificare i pesi delle connessioni tra i due neuroni. Nell'apprendimento non-supervisionato la rete impara senza che nessuno dall'esterno le dica nulla ed un esempio sono le reti di Kohonen.
Queste reti tengono conto non solo delle connessioni sinottiche tra neuroni ma anche dell'influenza che può avere un neurone vicino. Infatti nel caso biologico i neuroni che sono fisicamente vicini a neuroni attivi hanno legami più forti mentre quelli lontani legami più deboli. Una rete di Kohonen è costituita da una serie di neuroni di input e da un singolo strato bidimensionale di neuroni organizzati su una griglia posta su un piano. Ciascun neurone di input è connesso a tutti i neuroni della griglia e l'apprendimento è collegato alle interconnessioni tra neuroni vicini.
Nelle reti di Kohonen si confronta un pattern di input ed il vettore dei pesi :il neurone con il vettore dei pesi più vicino al pattern di input viene selezionato (J*). Questo nodo richiama il vettore di input e modifica il suo vettore di pesi in modo da allinearlo a quello dei pesi X. Si nota come vengono modificati anche i vettori dei pesi dei neuroni vicini a j* e questo perché la rete sta cercando di creare regioni costituite da un ampio set di valori attorno all'input

da cui apprende. Di conseguenza i vettori che sono spazialmente vicini ai valori di apprendimento saranno classificati correttamente anche se la rete non li ha mai visti. Questo dimostra la proprietà di generalizzazione della rete. Il valore di Nj* cioè il numero di neuroni vicini al neurone prescelto deve essere più grande possibile all'inizio e descrescere poi lentamente all'aumentare dei cicli di apprendimento.

3.4 Alberi decisionali

E' una tecnica di segmentazione che permette di costruire un grafo detto albero o dendrogramma. Dato un insieme n di osservazioni su una variabile dipendente (obiettivo) y e un certo numero di variabili esplicative (predittori) X, lo scopo dell'analisi degli alberi di decisione è quello di esplorare le relazioni tra le variabili mediante la suddivisione progressiva del campione iniziale in gruppi via via più omogenei al loro interno rispetto alla variabile dipendente detta anche "obiettivo" dell'analisi. Al primo passo, il campione di n unità viene diviso in due o più sottoinsiemi, caratterizzati dai valori assunti da due o più variabili esplicative ciascuno dei quali può essere ulteriormente suddiviso, fino a quando il processo viene interrotto in base ad una regola d'arresto.
Questa regola d'arresto deve tener presente il numero minimo di unità in un nodo, il numero massimo di nodi terminali, la minima disomogeneità del nodo, il massimo numero di passi del processo (e quindi il massimo numero di gruppi finali), la minima devianza del gruppo genitore (sotto la cui soglia un gruppo genitore si può già dire compatto e non più divisibile), una dimensione minima dei gruppi (sotto la cui soglia non si ha attendibilità del processo) e una soglia minima (fissata) di una funzione criterio della segmentazione che stabilisce se la suddivisione possa o meno aver luogo.
Questa funzione criterio stabilisce una misura della diversità tra i due o più gruppi figli generati dalla segmentazione di un nodo (più infatti questa misura è alta e più i sottogruppi generati sono omogenei tra loro).
Vi sono tre famiglie di questa funzione criterio $\Phi(s,t)$ dove t è il nodo ed s la segmentazione due delle quali sono le seguenti:

1) $\phi(s,t)=(\sum_r(|\tilde{y}_{tr}-\tilde{y}_t|^\lambda)*w_{tr})^{(1/\lambda)}$ con w_{tr} coefficiente di ponderazione

se $\lambda=2, w_{tr}=n_{tr}, \tilde{y}_t=\bar{y}_t$ si ottiene la devianza (rapportando la funzione al suo massimo cioè alla devianza del gruppo t si ha η^2 di Fisher)

1) $\phi(s,t)=-\sum_j p(j)*\log(p(j))+\sum_r\sum_j p(j/t_r)\log(p(j/t_r)$ che è l'entropia

L'analisi di segmentazione è applicabile qualunque sia la natura della variabile dipendente: quantitativa, qualitativa o dicotomica (quando la variabile è qualitativa si parla di alberi di classificazione quando è quantitativa di alberi di regressione).Le variabili esplicative possono essere anch'esse di

qualsiasi scala. E' però da tener presente che nel caso di variabili dipendenti qualitative si utilizza il caso univariato con una variabile dipendente alla volta, nel caso di variabili dipendenti quantitative si utilizza il caso multivariato che permette di considerare più variabili dipendenti contemporaneamente mentre nel caso dicotomico lo si assimila o al caso qualitativo o a quello quantitativo ma sono oggetto di ricerca nuovi criteri.
Si ha partizione binaria se si considerano partizioni a due vie, ternaria se si hanno partizioni a tre vie e segmentazione multipla se si hanno partizioni a k vie. Un ovvio vantaggio della segmentazione multipla è quello di costruire alberi meno profondi e più sintetici che risultano più facilmente leggibili e interpretabili, ma il problema sono i tempi di calcolo.
La ricerca della migliore segmentazione a volte è difficile perché il numero delle segmentazioni da esaminare aumenta molto velocemente con il numero dei predittori considerati.
E' inoltre da tener presente che la numerosità del campione deve essere sufficientemente elevata perché il frazionamento in sottogruppi può portare a stime dovute solo al caso.
Gli algoritmi oggi utilizzati comprendono CHAID (Chisquared Automatic Interaction Detection) per la segmentazione multipla di una variabile quantitativa, CART (Classification and Regression Trees) e C4.5 che generalizzano alcuni aspetti dell'approccio alla segmentazione e propongono nuove procedure per la costruzione e valutazione dell'albero.
Un esempio di un albero di decisione con segmentazione binaria, una variabile dipendente (la specie di Iris classificata nelle tre categorie: Setosa, Versicolor e Virginica) e due variabili esplicative lunghezza e larghezza del petalo è il seguente:

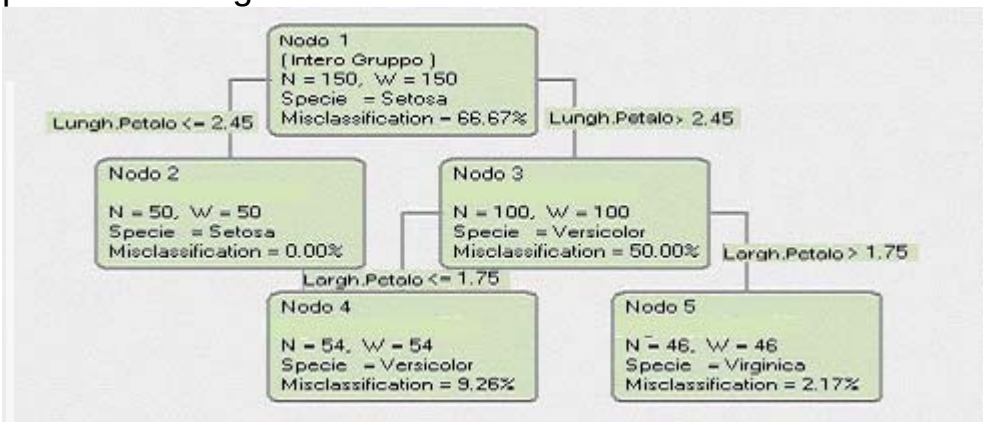


Fig. 3.4
I rettangoli mostrati nell'albero sono chiamati "nodi" e rappresentano i gruppi di unità ai diversi stadi del processo di segmentazione, i rami rappresentano le condizioni che hanno determinato le suddivisioni e le foglie sono i nodi terminali per i quali non è ritenuta utile alcuna suddivisione.
Il numero del nodo è mostrato in cima al rettangolo. La linea "N = nn" mostra la dimensione campionaria del nodo. La divisione viene selezionata per costruire un albero che può essere usato per predirre il valore di una variabile

obiettivo. Per ciascuna divisione, due decisioni vengono prese: quale variabile predittore usare per la divisione e quale set di valori della variabile predittore va nel nodo di sinistra e quale set va nel nodo di destra; questo è chiamato il punto di divisione.
Se la variabile di divisione è continua i valori che vanno a sinistra e a destra dei nodi figli saranno mostrati come valori minori o più grandi del valore del punto di divisione(che è 2.45 in questo esempio).

3.5 Analisi Cluster

La Cluster analisi è un insieme di procedure o algoritmi che si prefiggono di classificare o raggruppare individui in classi tali che: gli individui all'interno di una classe siano molto simili e ogni classe sia relativamente distinta dalle altre.
Fondamentalmente esistono due diversi tipi di algoritmi di classificazione: quelli gerarchici e quelli non gerarchici.
In generale, gli algoritmi gerarchici procedono alla creazione di gruppi di osservazioni attraverso unioni o divisioni (Gong e Richman 1995) e sono divisi in due principali categorie: agglomerativi e scissori.
Metodi Agglomerativi : questi metodi iniziano con ciascuna osservazione rappresentante un singolo cluster. A ciascun passo, due cluster vengono uniti fino al passo finale, dove rimane solo un cluster . Il ricercatore deve decidere a quale passo fermare la clasterizzazione, e cosi', quanti cluster conservare.
Metodi Scissori : i metodi di questo tipo lavorano in una maniera diversa dei metodi agglomerativi. Inizialmente, tutte le osservazioni iniziano in un grande

cluster. A ciascun passo, un cluster viene diviso fino a quando ciascuna unità costituisce un cluster. Anche qui, il ricercatore deve decidere il numero di cluster da conservare.

Misure di distanza: nel caso in cui sulle n unità da classificare si siano rilevate p variabili misurabili, un metodo per misurare la distanza tra le entità h e k si basa sulla **distanza euclidea**:

$$d(h,k)=\sum_{i=1}^{p} (x(h,i)-x(k,i))^2*w(i))^(1/2) \quad \text{dove } h,k=1,2,..,n$$

Per meglio capire questo la distanza tra due unità h e k in un sistema di assi cartesiani è l'ipotenusa che collega i due punti:

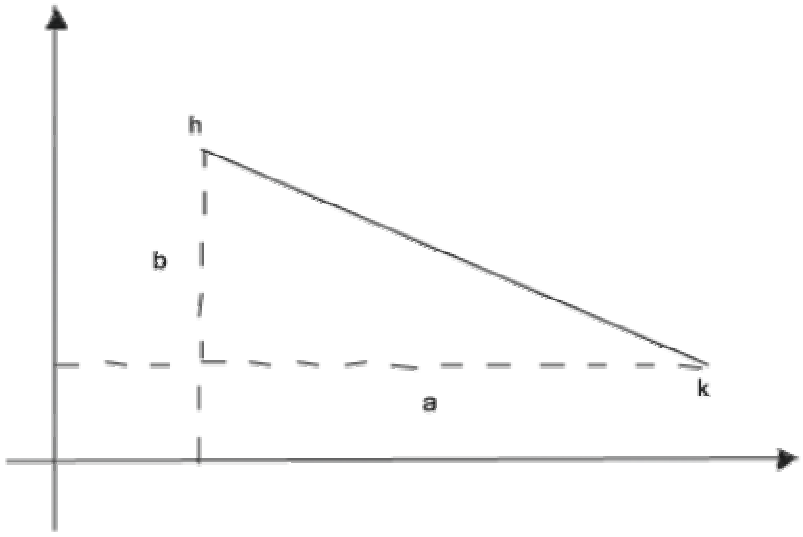


Fig. 3.5

Esistono comunque anche altri metodi per calcolare la distanza quale la **distanza media assoluta**

$$d(h,k)=\sum_{i=1}^{p}|x(h,i)-x(k,i)|*W(i) \quad h,k=1,\ldots,n$$

e la **distanza di Lagrange-Tchebychev**

$$d(h,k)=\sum_{i=1}^{p}\mathrm{Max}|x(h,i)-x(k,i)| \quad h,k=1,..,n$$

Metodi agglomerativi

1)Metodo del legame singolo: questo metodo misura la distanza tra cluster mediante la distanza tra i due più vicini punti all'interno dei cluster. Esso, insieme al legame completo rappresenta la più semplice tecnica di cluster .

Questo metodo viene illustrato in figure 3.5. Mentre è semplice da impiegare, questo metodo come Wilks (1995) descrive è problematico, in quanto vengono creati cluster molto grandi e non rappresentativi, a causa della vicinanza di punti a un "lato" di un cluster. Anche se la maggioranza dei punti sono lontani da ciascun altro, ne basta solo uno, vicino alla coppia di punti per causare l'aggregazione tra i due cluster. Questo metodo non è perciò molto famoso.

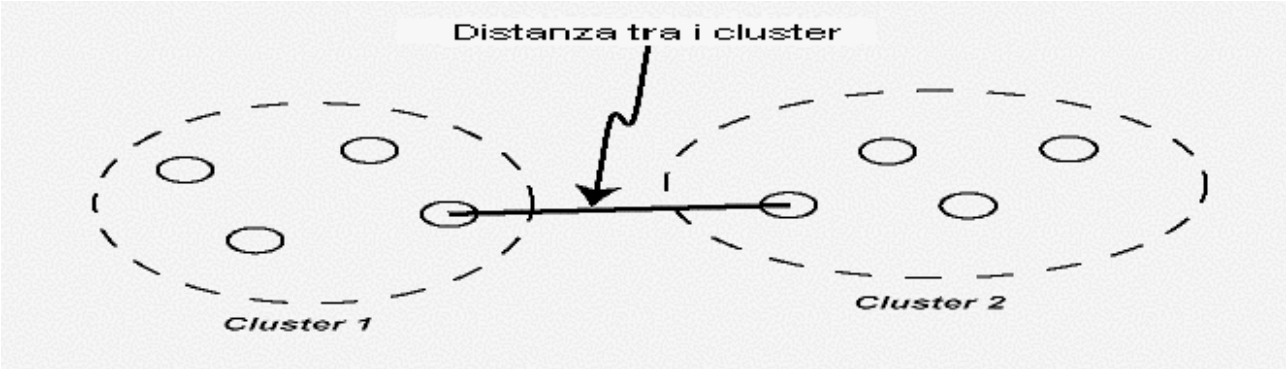


Fig. 3.5

2)Metodo del legame completo: Questo metodo lavora in maniera diversa rispetto al metodo del legame singolo. In questo metodo, la distanza tra cluster viene definita come la distanza tra la coppia di punti più lontana all'interno dei due cluster. Questo viene illustrato in figura 3.6. Mentre inizialmente questo metodo appare più attraente del legame singolo, Wilks (1995) indica che questo metodo crea un grande numero di clusters più piccoli che è rappresentativamente dovuto alla pessima qualità della misura di distanza.

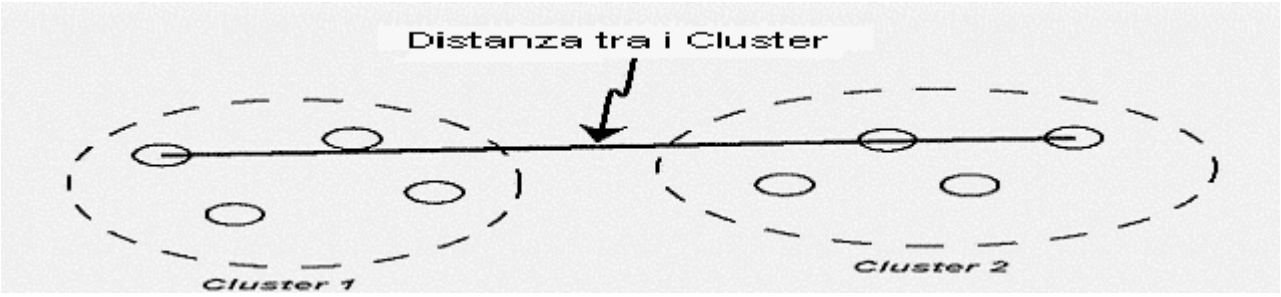


Fig. 3.6

3)Metodo del Centroide: questo metodo è un compromesso tra legame singolo e legame completo. In questo metodo, la posizione del cluster medio (es centroide) viene determinata con la posizione media di tutti i punti all'interno di quel cluster. A questo punto, la distanza tra cluster viene definita come la distanza tra la coppia di cluster centroidi. Una illustrazione di questo metodo viene data in figura 3.7. Sia nel singolo che nel legame completo, la distanza tra cluster da unire deve incrementare a ciascun passo di clasterizzazione (se non succede questo, vorrà dire che a un dato passo del processo, si aggregherà una coppia di cluster diversa da quella della coppia di cluster più vicina). Questa è una proprietà del fatto che nell'ambito del processo ciascuna osservazione rimane fissata nello spazio p-dimensionale. In questo metodo, invece, un cluster "assorbe" i membri da un'altro cluster e il suo centroide quasi certamente cambia di posizione. In conseguenza a questo, è possibile che la distanza tra cluster da unire decrementi con l'aumentare del numero di passi e questo rende i risultati del metodo centroide difficili da esaminare (il metodo non è perciò indicato).

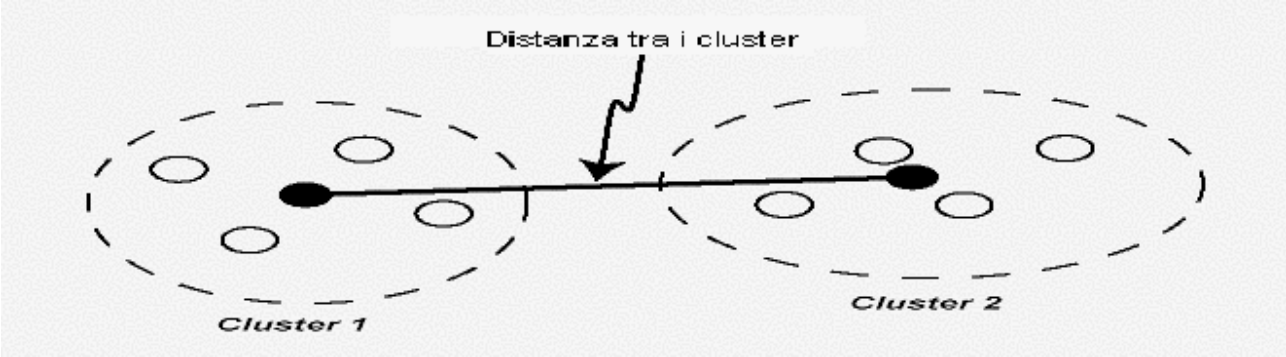


Fig. 3.7

4)Metodo della media di gruppo: questo metodo è un altro compromesso tra il legame singolo e quello completo. In esso, la distanza tra due cluster viene definita come la distanza media tra tutte le possibili coppie di punti all'interno dei due cluster. In questo modo, viene calcolata una "distanza media" e a ciscun passo vengono aggregati i due cluster con la più piccola distanza media tra i loro punti. Questo viene illustrato in figura 3.8. Poichè questo metodo non usa i centroidi, il problema del decremento della distanza di cluster da unire all'aumentare dei passi del processo non avviene. Una variazione a questo metodo è l'aggiunta di un peso che indica il numero di membri di un cluster. In questa maniera, vengono più probabilmente aggregati i più grandi che i più piccoli cluster.

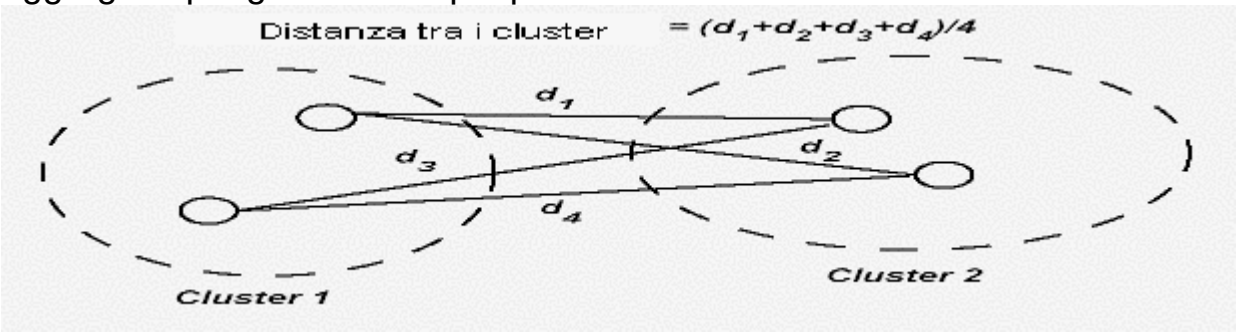


Fig. 3.8

5)Metodo di Ward(1963): in questo metodo , la scelta della coppia di cluster da aggregare si basa sulla minimizzazione della devianza tra i centroidi dei possibili gruppi. Wilks (1995) indica che questo metodo tende a creare cluster della stessa dimensione. La devianza ha un minimo pari a 0 quando tutte le entità sono isolate e un massimo pari alla somma delle devianze delle variabili di classificazione quando tutte le entità fanno parte di un singolo cluster. Ad ogni passo del processo vi sono indicatori statistici che indicano il numero ottimale di cluster da avere. Un esempio è l'indicatore R^2 che indica il rapporto tra la variabilità tra i cluster e la variabilità totale, dove la variabilità tra i cluster misura l'eterogeneità tra un gruppo e l'altro (più è alta e più differenziati sono i gruppi).

Altri indicatori sono l'R^2 semiparziale che è un peggioramento dell'R^2 dovuto all'agglomeramento del passo precedente, l'R^2 stabilito ad ogni passo, lo pseudo F Statistic che misura il grado di separazione tra i cluster ad

ogni livello gerarchico e lo pseudo t^2 Statistic che misura il grado di separazione tra gli ultimi due cluster aggregati.
Si arresta il processo di aggregazione al passo precedente quando il primo e il quarto indicatore assumono valori elevati o quando il terzo ed il secondo peggiorano bruscamente.

Metodi gerarchici scissori

Tra questi metodi vi è il metodo basato sulla **distanza tra i centroidi** ed il **metodo delle K Medie**.
-Metodo delle K Medie: questo metodo è molto differente da qualsiasi altro metodo menzionato sopra. Dopo la determinazione del numero di cluster che si vuole avere in un dataset con questo numero si indica anche il numero di "punti seme" che si vuole nel dominio. Il punto seme agisce come un cluster centroide, ed a questo livello è unico nel cluster. Nel passo successivo si piazzano le osservazioni più vicine al “punto seme” nel cluster. Si usa per esempio la distanza euclidea per determinare la misura di vicinanza tra le osservazioni. Successivamente, vengono ricalcolate le locazioni dei cluster centroidi. Poichè i centroidi si muovono, può essere possibile che certe osservazioni siano più vicine a un cluster centroide diverso dal cluster in cui sono presenti quelle osservazioni. Questo meccanismo agisce come correttore del precedente. I centroidi vengono aggiornati come nuovi membri che lasciano o vanno ad aggiungersi al cluster. Questa correzione è fatta fino a quando nessuna singola osservazione cambia cluster nel meccanismo. La figura 3.10 fa vedere questo processo.

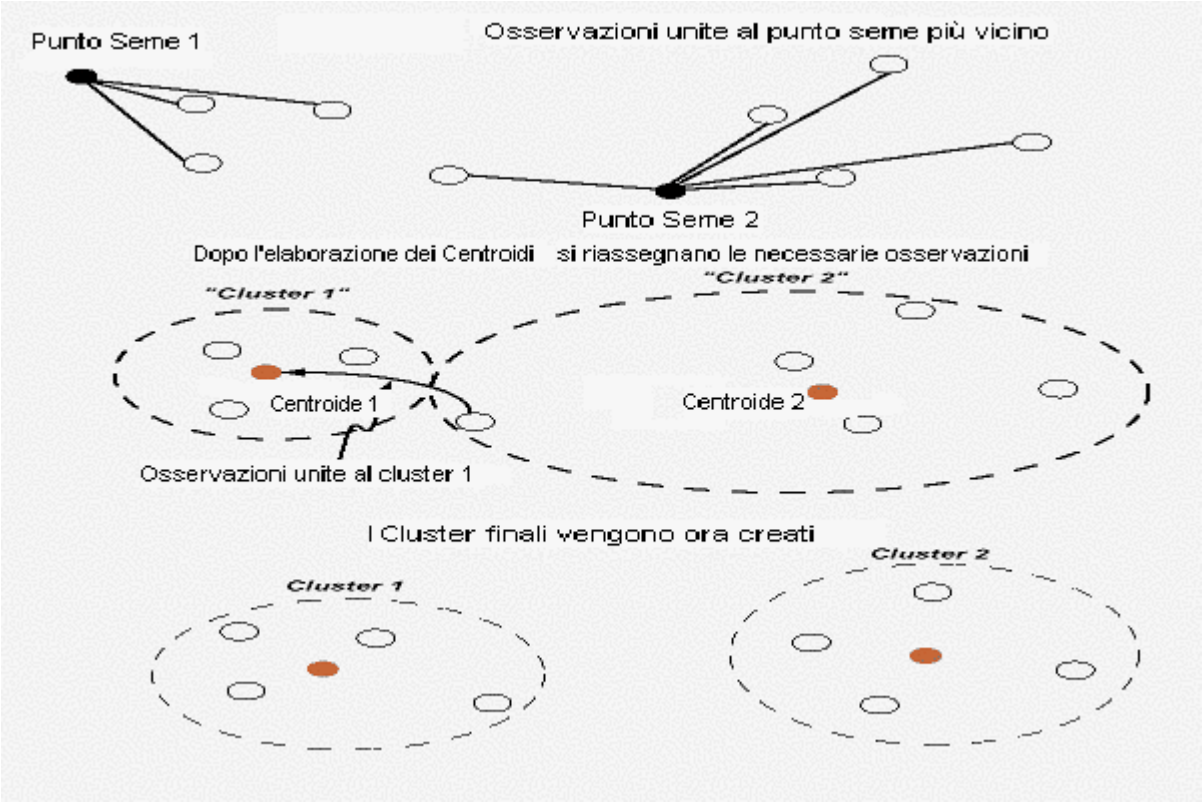


Fig. 3.10 I punti neri rappresentano i “punti seme”. I punti rossi rappresentano i centroidi dei cluster. I cerchi tratteggianti rappresentano "domini" di cluster durante quel passo di processo.

Metodi di analisi dei cluster non Gerarchici: questi metodi lavorano inizialmente dividendo le osservazioni in un dato numero di cluster e risistemando poi queste in differenti cluster fino a quando non si raggiunge lo scopo. La differenza principale tra questi metodi e i metodi gerarchici è la richiesta del numero di cluster prima che inizi il processo.Le tecniche non gerarchiche si suddividono in tecniche non gerarchiche **con sovrapposizione** e tecniche non gerarchiche che **generano partizioni**. Delle prime fanno parte la ricerca di insiemi sfuocati, l’analisi di miscugli di distribuzione e l’analisi fattoriale Q. Nelle seconde si seguono tre passi: determinazione a priori del numero di cluster, assegnazione delle entità ai cluster individuati prima e assegnazione delle entità a cluster diversi da quelli di partenza ottimizzando una funzione obiettivo. A questo punto si generano in automatico dei centroidi casuali e in base alla minima distanza euclidea dai centroidi si assegnano le unità ai diversi cluster. Si procede facendo deviare le unità dal cluster meno omogeneo ai due cluster più vicini e si ricalcolano poi i centroidi procedendo così finché gli spostamenti dei centroidi diventano irrilevanti.

3.6 ANALISI FATTORIALE

L'analisi fattoriale è una tecnica statistica che permette di spiegare la correlazione tra le variabili osservate in un numero ridotto di fattori o variabili "latenti" (in quanto non si possono osservare direttamente) ottenuti come combinazione lineare delle variabili originarie con una perdita minima di informazione.
Si supponga di avere osservato un insieme p di variabili quantitative presso n unità statistiche con n abbastanza grande rispetto a p. Si definisca l'elemento X_{hj} che denota il valore della variabile x_j osservato presso l'unità h e di aver standardizzato le osservazioni della matrice **X** come variabili di media 0 e varianza 1.
Il modello che ne deriva si esprime con l'equazione:

$$x_j=a_{j1}*f_1+a_{j2}*f_2+\ldots+a_{jq}*f_q+u_j*c_j \quad (j=1,2,..p)$$

dove con f_i si denota il fattore comune i-esimo e a_{ij} è il coefficiente che lega il fattore f_i alla variabile x_j detto anche peso fattoriale, c_j è il fattore specifico di x_j e u_j un suo coefficiente.
Nel modello fattoriale si ha: correlazione(f_i,f_j)=0 per ogni i diverso da j, correlazione(c_i,c_j)=0 per ogni i diverso da j,correlazione (c_i,f_j)=0 per ogni i,j.
Il fattore f_i si dice comune perché appartiene a tutte le p espressioni e se tutti gli a_{ij} sono non nulli per ogni i si dice anche generale.
I coefficienti a_{ij} coincidono con i coefficienti di correlazione tra fattori e variabili e più è alto il coefficiente più quella variabile è determinante per quel fattore.
In particolare :

$$f_i=w_{1i}*x_1+w_{2i}*x_2+..+w_{pi}*x_p$$

dove w_{ji} è il coefficiente fattoriale della variabile x_j nella combinazione f_i.

Il modello matematico di analisi fattoriale si basa sull'idea che il contenuto informativo di ciascuna variabile è dato da:
Var(x_i)=Informazione condivisa+Informazione specifica.

3.6.1 ANALISI DELLE COMPONENTI PRINCIPALI

L'analisi delle componenti principali è un metodo di analisi statistica multivariata che può essere derivata con molteplici approcci che conducono ad un unico algoritmo di decomposizione in autovalori/autovettori
L'analisi delle componenti principali è una trasformazione matematica di un insieme di variabili originarie X_i (i=1,..,r) dipendenti tra loro in un nuovo insieme di variabili Z_i (i=1,..,r) dette componenti principali la cui caratteristica è l'indipendenza reciproca. Sia,inoltre, S=s_{ij} (i,j=1,..,p) la relativa matrice di varianze-covarianze.

L'obiettivo è quello di trovare r trasformazioni lineari Z_i delle variabili originarie:

$Z_1=w_{11}*X_1+w_{21}*X_2+\ldots+w_{p1}*X_p$
$Z_2=w_{12}*X_1+w_{22}*X_2+\ldots.+w_{p2}*X_p$
.
.
$Z_r=w_{1r}*X_1+W_{2r}*X_2+\ldots. +w_{pr}*X_p$

La matrice S è data quindi da **X'X**/(n-1).
Si predispone la funzione l=**w**'**S**w-Ω(**w**'**w**-1) dove Ω è un moltiplicatore di Lagrange: essendo **Z**=**Xw** si ha l=**Z'Z**/(n-1)-Ω(**w**'**w**).
Imposto w'w=1 come funzione di normalizzazione del vettore w si deve massimizzare la funzione l e per far questo la si deriva rispetto a w e la si uguaglia a zero.
Si ha cioè δ(l)/ δ(w')=2**S**w-2Ω**w**=0 e quindi
(*)(**S**-Ω **I**)**w**=**0** dove **I** è la matrice identità di ordine p e **0** è un vettore colonna di zeri.
L'equazione (*) ha soluzione non nulla se e solo se il determinante di
S-Ω **I** è nullo la cui soluzione comporta r valori Ω detti autovalori (Ω_1>= Ω_2>=,.., >=Ω_r).
Inserendo Ω_1 nella (*) si ottiene (**S**-Ω_1 **I**)**w**$_1$ da cui si ricavano i coefficienti del vettore **w**$_1$ detti coefficienti della prima componente principale Z_1.
Si procede ora calcolando la matrice **S*** detta matrice di varianze-covarianze residua dove **S***=**S**-Ω_1**w**$_1$**w**$_1$'. A questo punto si massimizza la funzione l* sotto il vincolo **w**$_1$'**w**$_1$=1 ottenendo (**S***-Ω_2**I**)**w**$_2$=0 da cui si ricava il vettore **w**$_2$ dei coefficienti della seconda componente principale Z_2 e cosi' via fino all'ultima r-esima.
La prima componente principale è dunque la combinazione lineare che estrae il massimo di variabilità dalla matrice **S,** la seconda la combinazione **non** correlata con la prima che estrae il massimo della variabilità residua.
Si ha naturalmente che due componenti qualsiasi Z_i e Z_j sono indipendenti se i rispettivi vettori **w**$_i$**w**$_j$ sono ortogonali cioè **w**$_i$**w**$_j$=0 e che l'autovalore Ω_i rappresenta la varianza dell'i-esima componente principale. Inoltre la somma degli autovalori è uguale alla somma delle varianze delle variabili osservate.
Se invece della matrice **S** (matrice di variaze–covarianze) si utilizza la matrice **R** di correlazione dove per correlazione tra due variabili x e y si intende r=Cov(xy)/Var(x)Var(y), la somma degli autovalori è pari a p. Il prodotto degli autovalori è pari al determinante della matrice **S.**

Esempio:
Supponiamo di ottenere 12 autovalori dove Ω_1 =2.321, Ω_2 =2.02, Ω_3 =1.321, Ω_4 =1.001 e Ω_5=0.81,.. Ω_{12} =0.321.

Avremo quindi che la varianza totale è pari alla somma degli autovalori mentre la varianza spiegata dal primo e secondo fattore saranno:
2.321/12=0.1934 e 2.02/12=0.1683. Cumulativamente i primi 4 fattori spiegheranno il 55.52% della variabilità totale del fenomeno.
Una percentuale del 75% si considera un traguardo ma spesso si tollerano anche percentuali inferiori (60%) nel caso in cui vi siano molte variabili osservate (infatti con l'aumento del numero di variabili aumenta la variabilità estranea ai fattori primari).
Se l'analisi è svolta su una matrice di correlazione un metodo per estrarre il numero di fattori è quello di considerare gli autovalori>=1.

3.6.2 ROTAZIONE DEGLI ASSI

Abbiamo visto prima nell'analisi fattoriale che $x_j=a_{j1}f_1+..+a_{jq}F_q+u_jc_j$ con j=1,..,p.
Una rotazione dei fattori o degli assi ortogonali comporta a ridurre i coefficienti a_{ij} dell'analisi fattoriale in modo che quelli meno significativi alla prima analisi diventino ancora più piccoli (vicini a zero) e quelli più significativi più grandi (vicini a 1). I criteri di rotazione possono essere ortogonali o obliqui:
Tre criteri di rotazione ortogonali sono: varimax,quartimax e equamax.
Varimax: in questo criterio si tende a minimizzare il numero di variabili con cui ciascun fattore ha coefficienti di correlazione elevati. Tale criterio è raccomandabile se si vuole ottenere una netta separazione tra i fattori e se la rotazione è effettuata senza precisi criteri di riferimento. Non è raccomandabile se è estratto un solo fattore generale sul quale la maggior parte delle variabili ha pesi elevati.
Quartimax: mentre il Variamax semplifica le righe dei pesi fattoriali a_{ij} il Quartimax semplica le righe di questa matrice (a_{ij}) mantenendo inalterata la comunanza delle variabili (dove la comunanza della variabile x_j è data dalla somma dei coefficienti $a_{1j}^2+...+a_{qj}^2$) e massimizzando la varianza del quadrato dei coefficienti a_{ij} per l'insieme delle righe.
Tale criterio è adatto per identificare i fattori che governano la variabilità dei caratteri osservati e da risultati migliori del Varimax se si vuole semplificare il primo fattore estratto che tende ad essere un fattore generale.
Equamax:è un compromesso tra quartimax e varimax in quanto opera simultaneamente sulle righe e sulle colonne della matrice a_{ij}. Tale metodo non si adatta a strutture semplici.
Tra le rotazioni oblique abbiamo Promax che parte con una rotazione Varimax dei coefficienti a_{ij} e poi applica una procedura di aggiustamento dei pesi ruotati che incrementi i coefficienti a_{ij} già grandi e riduca quelli già piccoli (l'angolo degli assi viene cioè variato fino ad ottenere la soluzione ottimale).

La matrice di coefficienti a_{ij} dopo una rotazione obliqua è una matrice di coefficienti di correlazione parziale tra la variabile x_j e un fattore f_i al netto di tutti gli altri fattori.

4 TEXT MINING E WEB MINING

4.1 DEFINIZIONE DI TEXT MINING

Come il Data Mining è un processo di KDD (Knowledge Discovery in Database), il Text Mining è un processo di KDT (Knowledge Discovery in Text Database) in quanto mentre il primo scopre l'informazione nascosta da dati strutturati di grosse dimensioni, il secondo esegue quest'analisi su dati non strutturati o semi strutturati quali sono i documenti testuali servendosi per questo di un meccanismo intermedio (Intermediary Form) generato o da tecniche di Information Extraction (IE) che creano un database strutturato o da tecniche di Information Retrieval (IR) che creano una matrice di associazione "termini per documenti" su cui poi vengono effettuate (sul database o sulla matrice) le analisi di Data Mining per cercare le informazioni di interesse. In pratica il Text Mining è costituito da una serie di algoritmi che trasformano il testo in dati strutturati su cui viene poi effettuato il Data Mining.
L'Information Retrieval (IR) è quel processo che fornisce risposte (recupero di documenti) alle richieste formulate dagli utenti.
Voglio soffermarmi un attimo a parlare dell' Information Extraction (IE): mediante queste tecniche viene suddiviso il testo sorgente in un insieme di unità comprendenti parole, numeri, segni di punteggiatura dalle quali vengono eliminate quelle che fanno riferimento ad articoli, preposizioni, congiunzioni o voci verbali (che si presentano con alta frequenza non costituendo quindi importanza significativa per l'informazione). Si estraggono poi da queste unità filtrate le radici delle parole rimuovendo desinenze e suffissi e selezionando quelle unità che presentano nomi, trascurando quindi verbi, aggettivi, pronomi e avverbi. Alla fine si creano raggruppamenti di parole che costituiscono dei "concetti" e si applica un particolare modello che crea una matrice di associazione "termini-per-documenti" che si basa sulla logica che un termine è utile se la sua frequenza all'interno del documento è elevata ed presente su pochi documenti. Infatti la matrice gioca su due misure: la frequenza di un generico termine all'interno di un dato documento e l'inverso della frequenza che concorre un termine su tutti i documenti considerati. A questo punto la matrice viene sottoposta ad una sorta di riduzione di dimensionalità tramite un altro algoritmo e viene ottenuta una struttura su cui possono essere applicate le tecniche di Data Mining. Le tecniche di IR sono comunque più diffuse delle tecniche di IE.

4.2 DEFINIZIONE DI WEB MINING

Con la crescita del World Wide Web è divenuto necessario utilizzare strumenti automatici per trovare, estrarre o filtrare informazioni desiderate, tool quali il Web Mining che da un lato cerca informazioni e risorse da milioni di siti e database on-line (Web content Mining) e dall'altro scopre e analizza i pattern di accesso degli utenti da uno o più server Web o servizi on-line (Web usage Mining).

1)**Web content Mining**: l'etereogeneità e la mancanza di struttura che permea le sorgenti di informazioni del Web, quali ad esempio i documenti a ipertesto, ha generato difficoltà nel gestire, scoprire e organizzare l' informazione.

I tradizionali motori di ricerca quali Virgilio, Google, AltaVista hanno dato dei confort dal punto di vista della ricerca all'utente, ma non senz'altro sono la soluzione per avere un' informazione strutturata e dettagliata.

Recentemente comunque sono stati sviluppati strumenti quali **agenti Web Intelligenti** e **database approach** che forniscono informazione ad un più alto grado di qualità dal Web. I primi si suddividono in tre categorie:

1)Agenti di ricerca intelligenti che cercano informazione usando le caratteristiche di un particolare dominio(es. un profilo utente)

2)Informazione di filtraggio e categorizzazione che usano informazioni semantiche incorporate in strutture link per creare gerarchie cluster o documenti ipertesto oppure combinano tecniche gerarchiche di clustering per organizzare una collezione di documenti Web basati su informazioni concettuali.

3)Agenti Web Personalizzati che apprendono le sorgenti d'informazione Web dalle preferenze dell'utente.

I database approach si basano su tecniche di organizzazione e integrazione di dati eterogenei e semi-strutturati del Web in collezioni di risorse quali i database relazionali utilizzando query di database e tecniche di data mining per accedere e analizzare quest'informazione.

2)**Web user Mining**: essi riguardano la scoperta in automatico dei pattern d'accesso da uno o più server web. Questi strumenti si suddividono in **strumenti per la conoscenza dei pattern** e **strumenti per l'analisi dei pattern.**

Fra i primi (che scoprono i pattern) si collocano gli strumenti del Web Miner che scoprono regole di associazione e pattern sequenziali da logs di accesso al server, fra i secondi (che comprendono, visualizzano e interpretano questi pattern) si trova WebViz system che visualizza percorsi di pattern trasversali, o le tecniche di OLAP sui logs di accesso al server.

BIBLIOGRAFIA

Testi cartacei:

Susi Dulli,Vittorio Favero, *Modelli e strutture per il data warehousing*, Diade, 2000

Nicola Del Ciello, Susi Dulli, Alberto Saccardi, *Metodi di data mining per il customer relationship management*, FrancoAngeli, 2000

Luigi Fabbris, *Statistica multivariata analisi esplorativa dei dati,* McGraw-Hill, 1997

W.H. Inmon, *Building the data Warehouse*, John Wiley & sons, Inc., 1996

Jill Dychè, *e-Data,* Pearson Publication Company, 2000

Materiale dalla rete:

http://met.psu.edu/~arnottj/newclusterpage/Cluster-Analysis-Description.html

(1) http://www.goldnet.it/~daniele/tesi/node4.html

http://www.spss.it/solutions/datamine/techniques.htm

INDICE ANALITICO