Knowledge Discovery e Data Mining

Sommario

- Introduzione
  - Motivazioni ed applicazioni
- Il processo di KDD
  - Fasi e caratteristiche
- Le tecniche di DM
  - Classificazione e regressione
  - Scoperta di regole associative
  - Clustering
- La classificazione
  - Alberi di decisione
  - Classificatori instance-based

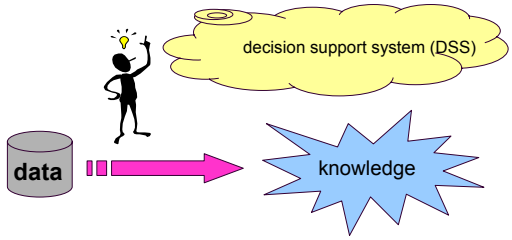
Motivazioni

- Problema della "crescita dei dati":
  - La disponibilità di tecnologie efficienti e poco costose per la raccolta e la gestione di dati porta ad accumulare un'incredibile quantità di dati
- Opportunità di trarre vantaggio competitivo dai dati
- Inadeguatezza dei sistemi di analisi tradizionali
  - "We are drowning in data, but starving for knowledge!"

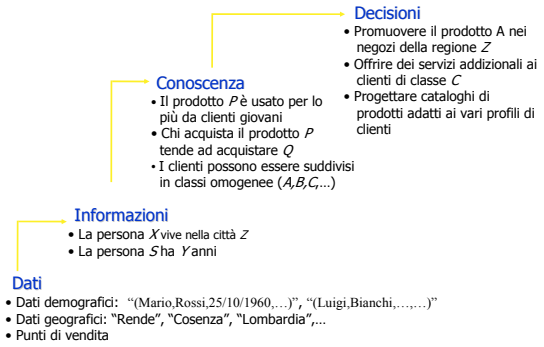
- Soluzioni: data warehousing e data mining:
  - On-line analytical processing
  - Estrazione di conoscenza interessante (regole, regolarità, patterns, vincoli) dai dati in grandi DB.

A cosa serve il Data Mining?

Accrescere la conoscenza su cui basare le decisioni



La catena del valore dell'informazione



Analisi dei dati ed estrazione di conoscenza

- KDD (Knowledge Discovery in Databases)
  - Processo di analisi dei dati finalizzato alla scoperta di nuova conoscenza, accurata e utile per
    - comprendere/modellare i fenomeni del mondo reale corrispondenti ai dati
    - supportare processi decisionali
- Data mining
  - È la fase chiave del processo di KDD
  - Scoperta automatica di pattern o di modelli a partire da un insieme di dati
    - *Pattern*: descrive regolarità (proprietà frequenti) dei dati
    - *Modello*: generalizza i dati e modella i fenomeni del mondo reale che li hanno generati

Data Mining

- riferito anche come:
  - Data dredging, Data harvesting, Data archeology
- un settore interdisciplinare:
  - Database e data warehousing
  - Data visualization
  - Statistica e machine learning
  - Sistemi esperti e knowledge acquisition

Motivazioni: perché ORA

- Abbondanza di informazioni utili
- Focus competitivo - Knowledge Management
- Sistemi di calcolo economici e potenti
- Fondamenti teorici/matematici robusti
  - machine learning & inferenza logica
  - statistica
  - database management systems

Estrazione della conoscenza: il processo di KDD

- Knowledge Discovery in Databases (KDD)
- Scoperta di conoscenza nelle basi di dati
  - utile per comprendere/modellare i fenomeni del mondo reale corrispondenti ai dati
- Processo non banale (semi-automatico) di estrazione di pattern
  - validi
  - nuovi
  - potenzialmente utili
  - comprensibili

} **INTERESSANTI**

Aree Applicative

- Analisi dei dati & supporto alle decisioni
  - Market analysis & management
  - Risk analysis & management
  - Fraud detection & management
- Altre Applicazioni:
  - Text mining (news group, documents) e Web analysis
  - Intelligent query answering

Applicazioni: Market Analysis & Management

- Quali sono le sorgenti di dati per l'analisi?
  - Transazioni di carte di credito, carte fedeltà, coupons di sconti, chiamate a clienti "accondiscendenti", studi su "stili di vita".
- Target marketing:
  - trova raggruppamenti che modellino gli acquirenti con stesse caratteristiche: interessi, abitudini di spesa, etc.
- Determina i pattern degli acquirenti nel tempo:
  - Conversione da un conto singolo a un conto congiunto: matrimonio, etc.
- Analisi incrociate
  - Associazioni/Correlazioni tra vendite di prodotti
  - Predizione basata sulle associazioni.

Applicazioni: Market Analysis & Management (2)

- Customer profiling
  - quali acquirenti comprano cosa (clustering o classificazione)
- Identificazione delle richieste dei clienti
  - identificare i migliori prodotti per acquirenti differenti
  - usa la predizione per capire quali fattori attraggono nuovi clienti
- Informazioni generalizzate
  - vari summary reports multidimensionali
  - informazioni statistiche aggregate (tendenze e variazioni)

Applicazioni: Risk Management

- Pianificazione finanziaria:
  - analisi e predizione dei flussi di cassa
  - valutazione dei profitti e analisi delle richieste contingenti
  - analisi di serie temporali (financial-ratio, trend analysis,...)
- Resource planning:
  - riassumi a confronte le risorse e le spese
- Competition:
  - monitora i vari competitori e le direzioni di mercato (CI: competitive intelligence).
  - Segmenta i clienti per classi di prezzi
  - Seleziona una strategia in un mercato altamente competitivo

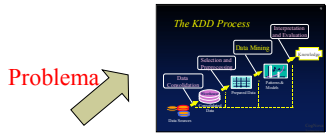
Applicazioni: Fraud Detection & Management

- Applicazioni:
  - monitoraggio della salute, prestiti, servizi di carte di credito, telecomunicazioni (frodi su carte telefoniche), etc.
- Approccio:
  - utilizza i dati storici per costruire un modello del comportamento fraudolento e utilizza il data mining per facilitare l'identificazione di istanze simili
- Esempi:
  - Assicurazioni auto: individua un gruppo di persone che creano (artificialmente) incidenti per guadagnare dalle assicurazioni
  - Money Laundering: individua movimenti di denaro sospetti
  - Assicurazioni mediche: individua i pazienti "professionali" dei medici
  - individuazione di frodi telefoniche

Altre Applicazioni

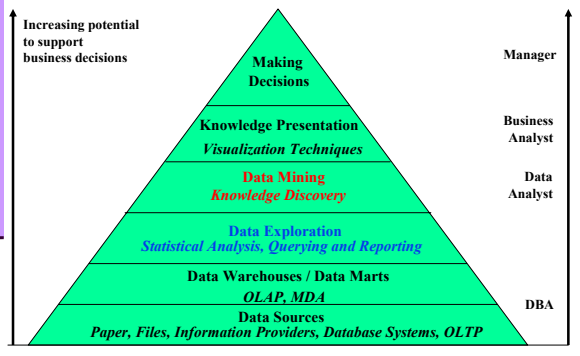
- Sport
  - IBM Advanced Scout hanno analizzato le statistiche della NBA
- Astronomia
  - Cal Tech e il Palomar Observatory hanno scoperto 22 nuove quasar con l'aiuto del data mining
- Internet Web Surf-Aid
  - IBM Surf-Aid applica algoritmi di data mining ai log di accesso delle pagine correlate alle informazioni commerciali, per scoprire le preferenze del cliente e indirizzare opportunamente il web-marketing

Il "Ciclo Virtuoso"

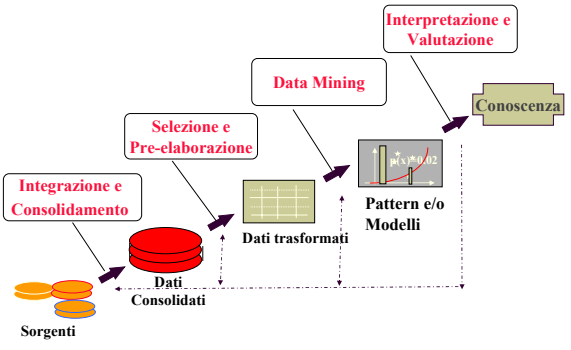


Problema → Identifica problema/ opportunità → Strategia → Misura effetti delle azioni → risultati → Agisci sulla base della conoscenza → conoscenza

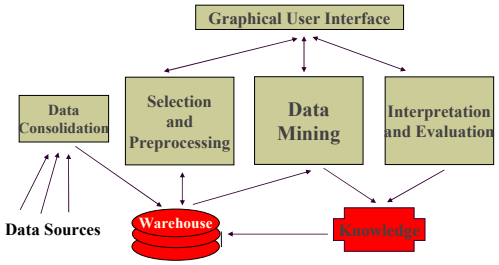
Data Mining & Business Intelligence



Il processo di KDD



Architettura di un sistema KDD



Consolidamento dei Dati

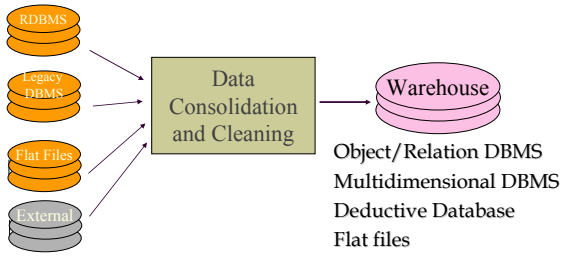
- Garbage in ⇨ Garbage out
  - La qualità dei risultati è in relazione diretta con la qualità dei dati
  - Il 50%-70% degli sforzi sul processo di KDD viene speso nella preparazione e il consolidamento dei dati

Consolidamento dei Dati

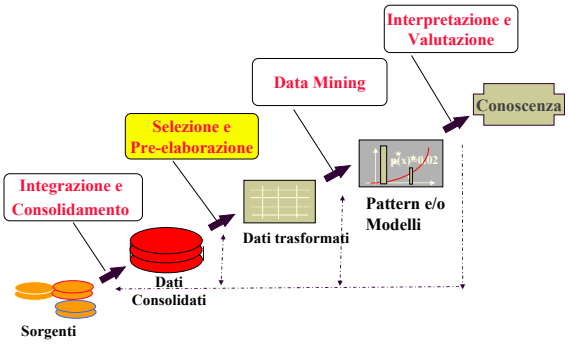
- Determinazione un insieme preliminare di attributi
- Consolidamento dei dati in un database
  - sorgenti interne ed esterne
- Eliminazione e/o stima dei valori nulli
- Rimozione di outlier (eccezioni ovvie, picchi)
- Determinazione della probabilità delle categorie e trattamento dei volume bias

Consolidamento dei Dati

Da sorgenti eterogenee a un data warehouse



Il processo di KDD



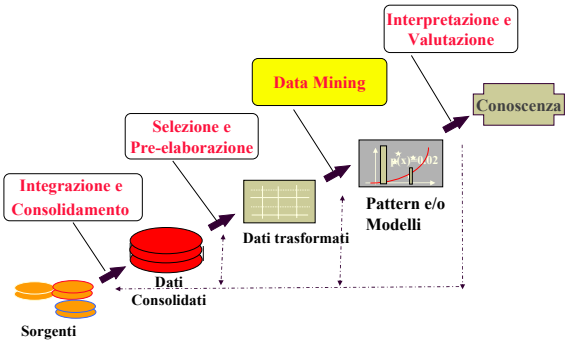
Selezione & Preprocessing

- Generazione di un insieme di esempi
  - scelta del metodo di sampling
  - analisi della complessità del campione
  - tratta le problematiche collegate al volume “bias”
- Riduzione delle dimensioni degli attributi
  - rimozione di attributi ridondanti e/o correlati
  - combinazione di attributi (somma, prodotto, differenza)
- Riduzione dei range di valori di attributi
  - raggruppamento di valori simbolici o discreti
  - discretizzazione dei valori continui

Selezione & Preprocessing: tool OLAP

- Può essere utile nella fase di selezione e preparazione dei dati
  - "esplorazione" dei dati, utile per individuare le caratteristiche rilevanti e suggerire i tipi di pattern/modelli che potrebbero essere scoperti
- Una visione multidimensionale dei dati
  - Aggregazioni per ogni intersezione di ogni dimensione.
- Analisi interattiva dei dati
  - Modellazione analitica: derivazione delle proporzioni, delle varianze, etc
  - previsione, trend analysis, e statistical analysis.
  - Calcola e visualizza i dati in 2D o 3D crosstabs, charts e grafi, con semplici operazioni di pivoting degli assi

Il Processo di KDD



Data Mining: un approccio di analisi induttiva

- Approccio deduttivo ("Verification-Driven")
  - Top-Down: analisi "passiva", atta a verificare se un certo modello (ipotesi) è coerente con i dati a disposizione
  - L'ipotesi o il modello sono formulati dall'utente sulla base della sua esperienza
- Approccio induttivo ("Discovery-Driven")
  - Bottom-Up: analisi "attiva", in cui i dati stessi suggeriscono possibili ipotesi sul significato del loro contenuto
  - Individuazione di fatti significativi, relazioni, tendenze, pattern, associazioni, eccezioni e anomalie, che sfuggono all'analisi manuale per la loro complessità

Approccio deduttivo

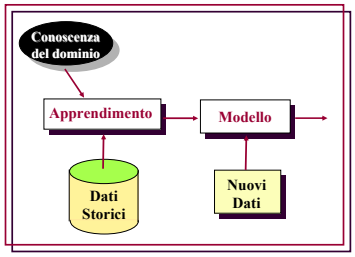
- "Query & Reporting"
- Foglio Elettronico
- Analisi multidimensionale
- Analisi statistica
  - Raccolta dei dati
    - Generazione di una ipotesi
    - Formulazione di una domanda
    - Interpretazione dei risultati e revisione dell'ipotesi

    ..... finchè non emerge un "pattern" interessante

Approccio induttivo (Data Mining)

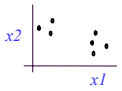
- Riconoscimento di Forme
- Intelligenza Artificiale
- Visualizzazione

Apprendimento automatico
simbolico
Reti neurali
Algoritmi Genetici
Reti Bayesiane

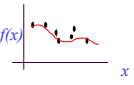


Data mining: task e metodi

- Esplorazione/Scoperta
  - scoperta di associazioni/correlazioni
  - clustering
    - algoritmi matematici, statistici, reti neurali, …
  - es. scoperta segmenti di mercato



- Predizione
  - uso di modelli di classificazione e regressione
    - matematici, statistici, reti neurali, algoritmi genetici, alberi e regole di decisione, …
  - es. prevedere volumi di vendita, o classe di cliente



- Descrizione
  - Scoperta di pattern/modelli facilmente interpretabili
    - regole/alberi di decisione, regole associative
  - es. descrivere clienti in base a dati demografici e acquisti

SE età > 35
E reddito < $35k
ALLORA...

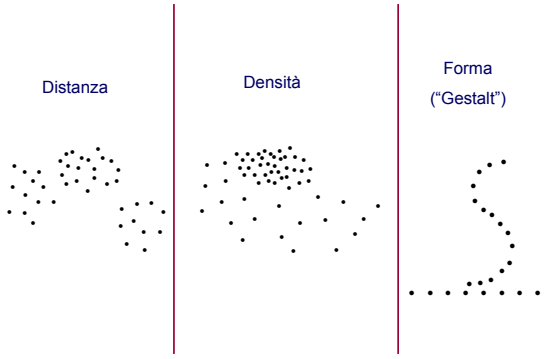
Data Clustering Analysis

- Clustering:
  - Partizionamento di un insieme di dati in un insieme di classi, chiamati cluster, tali che i membri di ogni classe condividano delle caratteristiche rilevanti
- Qualità dei cluster:
  - la similarità intra-classe é alta
  - la similarità inter-classe é bassa

3 categorie di Clustering

- Partizionali:
  - si enumerano le varie partizioni e quindi si valuta il partizionamento in base a qualche criterio.
  - K-means, K-medoids, etc.
- Agglomerativi-based:
  - si crea un agglomerato di dati utilizzando qualche criterio
  - Hierarchical Clustering
- Model-based:
  - Un modello viene ipotizzato per ogni cluster
  - a quel punto si trova qual'é la migliore costruzione che soddisfa il modello.
  - E.g., classificazione Bayesiana (AutoClass), Cobweb.

Clustering



Scoperta di Associazioni

- Scoperta di associazioni tra fatti, proprietà o valori di variabili ("Link analysis")
- La scoperta di associazioni tra variabili è solo il primo passo di analisi:
  - occorre cercare una spiegazione
- Esempio:
  - *Il 72% degli acquirenti di insalata verde, acquista anche qualche tipo di "dressing"*

Regole Associative

- Dati:
  - un database di transazioni
  - Ogni transazione é un insieme di item
- Obiettivo:
  - trovare tutte le regole che correlano la presenza di un insieme di item con un altro insieme di item
- Esempio:
  - 98% delle persone che comprano pannolini e cibo per bambini comprano anche birra.

Qualità di Regole Associative: supporto e confidenza

- Una regola deve avere una confidenza minimale, specificata dall'utente
  - 1 & 2 => 3 ha confidenza 90% se quando viene acquistato 1 e 2, nel 90% dei casi viene acquistato anche 3.
- Una regola deve avere un supporto minimale, specificato dall'utente
  - 1 & 2 => 3 dovrebbe valere in qualche percentuale minima delle transazioni per essere significativa

Regole Associative: Esempio

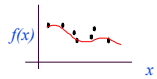
Transaction ID	Purchased Items
1	{1, 2, 3}
2	{1, 4}
3	{1, 3}
4	{2, 5, 6}

- Per un supporto minimo del 50%, e confidenza minima del 50%, troviamo le seguenti regole
  - 1 => 3 con 50% supporto e 66% confidenza
  - 3 => 1 con 50% supporto e 100% confidenza

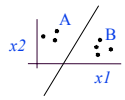
Modellazione Induttiva = apprendimento induttivo

Obiettivo: *Sviluppo un modello generale o un'ipotesi da esempio specifici*

- Approssimazione di funzione (curve fitting)

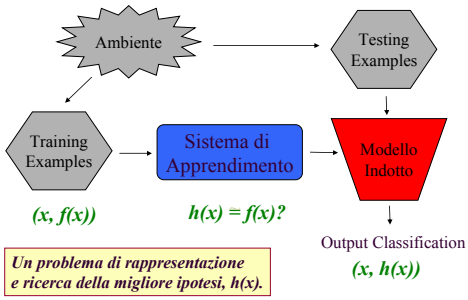


- Classificazione (concept learning, pattern recognition)



Classificazione e apprendimento induttivo

Framework di base per l'Apprendimento Induttivo:



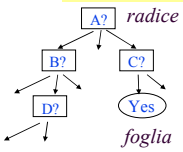
Un problema di rappresentazione e ricerca della migliore ipotesi, h(x).

Tecniche di apprendimento automatico

- Tecniche simboliche ("logiche")
  - Alberi di Decisione
  - Regole di Produzione
  - …
- Altri approcci
  - tecniche matematiche
  - tecniche statistiche
  - …

Alberi di Decisione

- Un grafo diretto aciclico
- I nodi sono:
  - Foglie - indicano la classe o distribuzione delle classi
  - Nodi di decisione - un test su un singolo attributo
    - un ramo ed un sottoalbero per ogni possibile risultato del test
- Classificazione:
  - navigazione da radice a una foglia, in base ai test



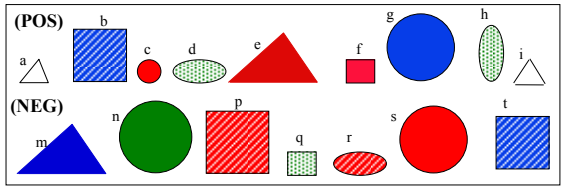
Alberi di Decisione : Esempio

Problema: indurre un albero che separi le classi di forme **(POS)** e **(NEG)** :

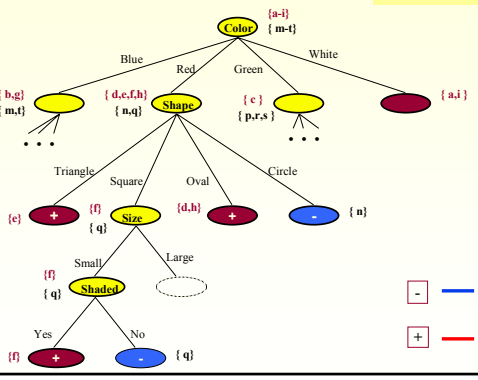
Attributi

Color = {Red, Blue, Green, White}
Shaded = {Yes, No}
Shape = {Square, Triangle, Circle, Oval}
Size = {Small, Large}

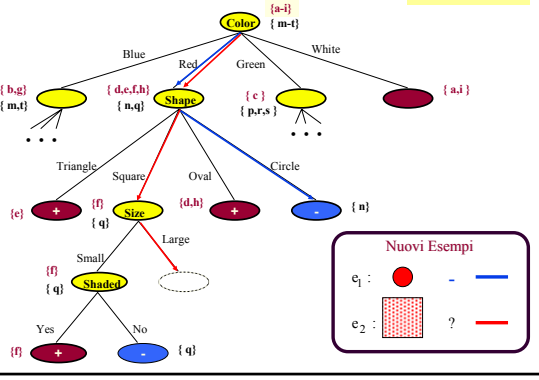
Esempi (training set):



Alberi di Decisione: Costruzione (apprendimento)



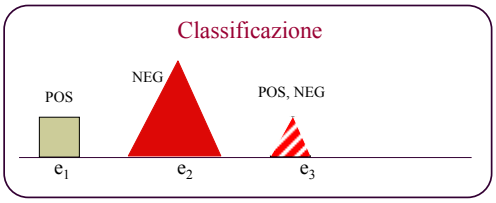
Alberi di Decisione: Uso (classificazione/predizione)



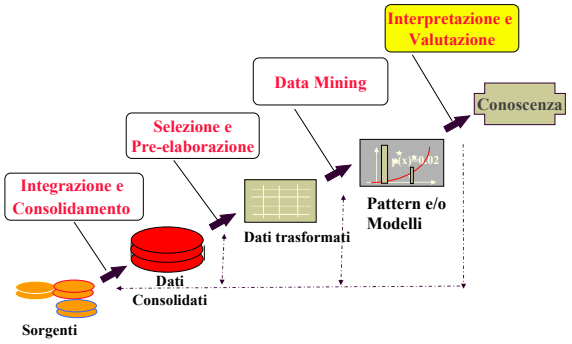
Regole di decisione

Regole espresse in forma logica:
(*Calcolo Proposizionale* o *Calcolo dei Predicati*)

(forma = quadrato or triangolo) and (dimensione = piccolo) => POS
(forma = triangolo) and (tratteggiato = SI) => NEG



Il Processo di KDD



Interpretazione e valutazione

- Interpretazione
  - Alberi e regole logiche possono essere interpretati facilmente
  - Risultati di clustering possono essere graficati/tabellati
  - In generale, tool di visualizzazione possono essere molto utili per interpretare i risultati
- Valutazione
  - Validazione statistica e test di significanza
  - Analisi qualitativa da parte di esperti del dominio
  - Test del modello su casi di esempio per valutarne la bontà (es., accuratezza)

Sono tutti interessanti i pattern scoperti?

- Un sistema di mining può generare migliaia di pattern
- Pattern interessanti:
  - Facilmente comprensibili
  - Validi su nuovi dati di test con un certo grado di certezza
  - Potenzialmente utili per processi decisionali
  - Nuovi o confermano ipotesi che si cercava di validare
- Misure Oggettive vs. Soggettive
  - *Oggettive*: basate su statistiche e struttura dei pattern (e.g., supporto, confidenza, lunghezza della descrizione, etc.)
  - *Soggettive*: basate su conoscenza pregressa e aspettative degli utenti (e.g., novità, etc.)

Completezza/Precisione di un processo di data mining

- Completezza: Trova tutti i pattern interessanti
  - E' possibile in un tempo ragionevole?
- Correttezza: Trova solo pattern interessanti
  - E' possibile "non commettere errori" e trovare solo pattern interessanti?
- Possibili approcci:
  - Si generano tutti i pattern che potrebbero essere interessanti e poi si eliminano quelli che non lo sono
  - Generare solo pattern interessanti (è possibile perdere in completezza)