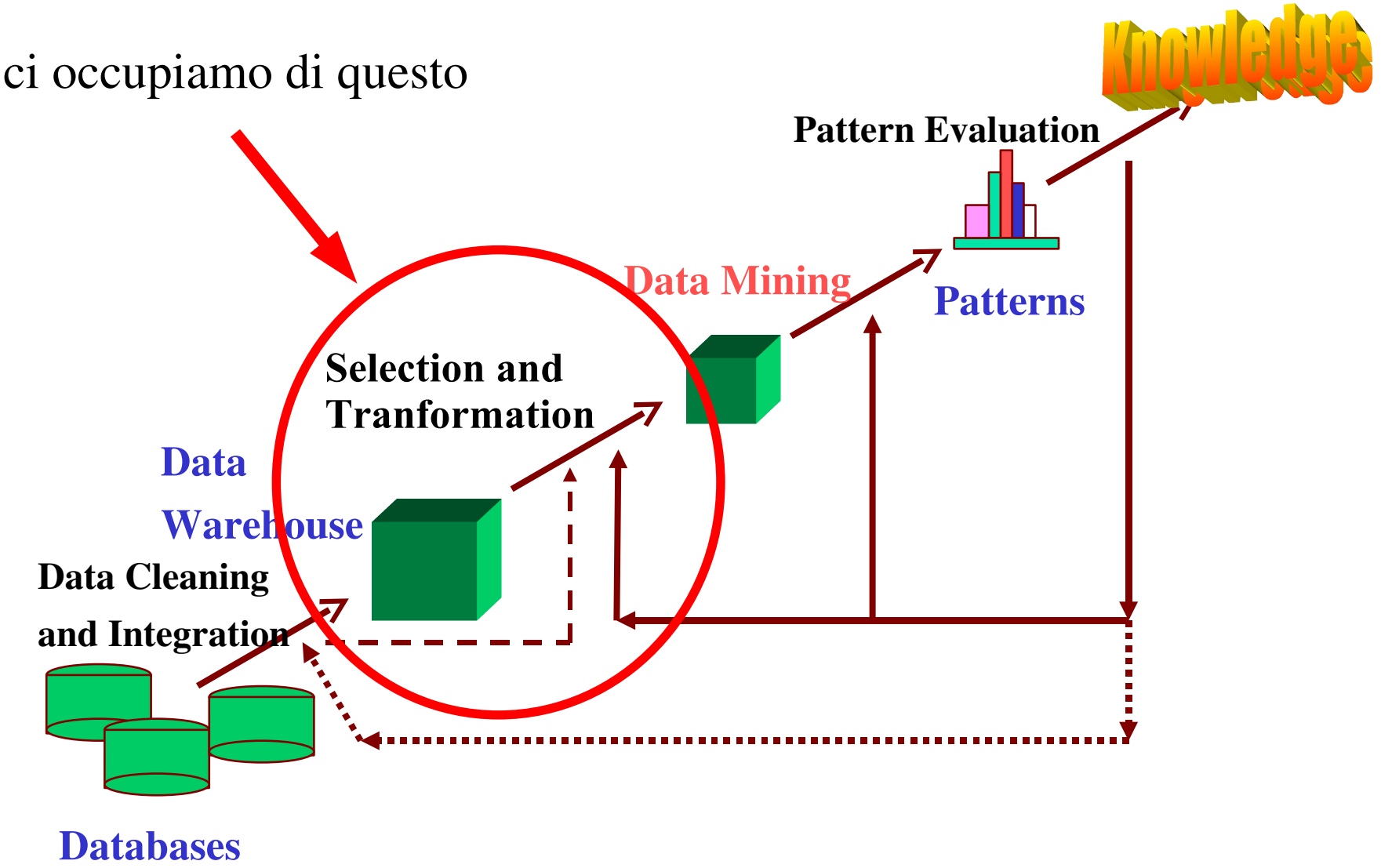
Preparazione dei dati

Gianluca Amato

Corso di Laurea Specialistica in Economia Informatica
Università “G. D'Annunzio” di Chieti-Pescara
ultima modifica: 12/05/08

Knowledge Discovery in Databases



Istanze e attributi

Input a un sistema di data mining

- Mentre nei sistemi OLAP la forma preferita di dati è il cubo multidimensionale, per i sistemi di data mining la visione tabellare è di solito più conveniente.
- L'input corrisponde essenzialmente a una tabella di un database relazionale:
  - ogni riga della tabella è una istanza (o esempio, o tupla)
  - ogni colonna è un attributo
- L'input è dunque un insieme di istanze, ognuna delle quali è un esempio indipendente dell'informazione che si vuole apprendere.

Un esempio di Input

	lun. sepalo	larg. sepalo	lun. petalo	larg. petalo	tipo
1	5,1	3,5	1,4	0,2	Iris setosa
2	4,9	3	1,4	0,2	Iris setosa
3	4,7	3,2	1,3	0,2	Iris setosa
4	4,6	3,1	1,5	0,2	Iris setosa
5	5	3,6	1,4	0,2	Iris setosa
..					
51	7	3,2	4,7	1,4	Iris versicolor
52	6,4	3,2	4,5	1,5	Iris versicolor
53	6,9	3,1	4,9	1,5	Iris versicolor
..					
103	7,1	3	5,9	2,1	Iris virginica
104	6,3	2,9	5,6	1,8	Iris virginica
105	6,5	3	5,8	2,2	Iris virginica

Attributi (1)

- Gli attributi possono essere distinti secondo il "livello di misura"
  - nominali
    - ogni valore è un simbolo distinto
    - l'unica operazione permessa è decidere se due valori sono uguali
    - ad esempio l'attributo tipo per il data set degli iris è un attributo nominale che assume tre possibili valori
  - ordinali
    - come gli attributi nominali, ma in più c'è un ordine
      - è possibile confrontare due valori co tutti gli operatori relazionali (<, >, = e derivati)
    - ad esempio, l'attributo temperatura può assumere i valori freddo, tiepido, caldo con freddo < tiepido < caldo.
      - spesso si usano i numeri interi per rappresentare i valori ordinali, dato che per essi è definito un ordinamento standard.
      - in questo caso, si potrebbe usare 0 per il freddo, 1 per il tiepido, 2 per il caso.
    - la distinzione tra attributi nominali ed ordinali non è sempre chiara.

Attributi (2)

- intervallo
  - assumono valori ordinati e ottenuti da precise unità di misura
  - esiste il concetto di distanza, per cui è possibile sottrarre due valori
  - le altre operazioni aritmetiche come divisione e prodotto non hanno senso, e non esiste un valore zero significativo.
  - ad esempio la temperatura, quando espressa in gradi Celsius o l'attributo anno.
- ratio
  - assumono valori ordinati e ottenuti da precise unità di misura, per cui esiste un valore zero ben definito.
  - tutte le operazioni aritmetiche hanno sesnso.
  - ad esempio la temperatura, quando espressa in gradi Kelvin, o la maggior parte delle misure fisiche come massa, lunghezza, etc..

Attributi (3)

- In pratica, la maggior parte delle volte gli algoritmi di data mining trattano solo due classi di attributi:
  - nominali (chiamati anche categoriali o discreti)
  - numerici: corrispondono ai tipi ordinale, intervallo o ratio a seconda del tipo di algoritmo
    - assumono un qualunque valore numerico
    - bisogna stare attenti al fatto che l'algoritmo non faccia delle operazioni che non hanno senso sul tipo di dato in questione

Metadati

- I sistemi di data mining possono usare altre informazioni oltre al tipo degli attributi:
  - informazioni dimensionali, in modo da non confrontare dati espressi con unità di misura diverse (cosa vuol dire che 3 Km è minore di 5 Litri?)
  - ordinamenti circolari: indicare se un attributo è soggetto a particolare circolarità dei dati
    - gli angoli vanno da 0 a 360° (o da 0 a 2π) e poi ricomnciano da 0.
    - ci si può riferire allo "stesso giorno nella prossima settimana" o alla "prossima domenica"
  - gerarchie di concetti: alcuni attributi possono essere trattati a vari livelli di dettaglio
- Tutte queste informazioni prendono il nome di metadati e consentono di aumentare l'efficienza del sistema di data mining.

Perché pre-elaborare i dati?

Pre-elaborazione dei dati (1)

- I dati nel mondo reale sono sporchi:
  - incompleti: manca il valore di alcuni attributi, o mancano del tutto alcuni attributi interessanti.
  - inaccurati: contengono valori errati o che si discostano sensibilmente da valori attesi.
    - Ad esempio, nel campo età di un impiegato si trova il valore di 120 anni.
  - inconsistenti: ad esempio, due filiali dello stesso negozio usano codici diversi per rappresentare la stessa merce venduta.
- Queste inesattezze non influivano sullo scopo iniziale per cui i dati sono stati raccolti, per cui vengono scoperte solo ora.
- GIGO: garbage in – garbage out
  - se i dati in input non sono di buona qualità, neanche le analisi basate su di questi lo possono essere!

Pre-elaborazione dei dati (2)

- Principali tecniche nella fase di pre-elaborazione dei dati:
  - data cleaning (pulitura dei dati)
    riempire i campi con i valori mancanti, “lisciare” i dati rumorosi, rimuovere i valori non realistici.
  - data integration (integrazione dei dati)
    integrare dati provenienti da database multipli risolvendo le inconsistenze.
  - data transformation (trasformazione dei dati)
    preparare i dati per l'uso con alcuni particolari algoritmi di analisi.
  - data reduction (riduzione dei dati)
    ridurre la mole dei dati in input, ma senza compromettere la validità delle analisi (campionamento, astrazione dei dati con le gerarchie di concetti, ...)

Data Cleaning

Data Cleaning

- Le attività eseguite durante il passo di data cleaning sono:
  - riempire gli attributi che hanno valori mancanti
  - identificare gli outliers (dati molto diversi dai valori attesi)
  - eliminare il rumore presente nei dati
  - correggere le inconsistenze
- Alcuni algoritmi di analisi hanno dei meccanismi per gestire dati con valori mancanti o con outliers.
  - essi operano però senza conoscenza del dominio applicativo
- I risultati migliori si ottengono con una pulizia a priori dei dati, con l'aiuto di esperti del dominio applicativo.

Dati mancanti (1)

- Varie ragioni per cui i dati mancano
  - malfunzionamento di qualche apparecchiatura.
  - dati inconsistenti con altri e quindi cancellati in una fase precedente.
  - dati non immessi.

- Mancanze casuali o no?
  - se un valore non è presente perché un determinato test non è stato eseguito in maniera deliberata, allora la presenza di un attributo mancante può veicolare una grossa mole di informazione.
    - le persone che studiano i database di natura medica hanno scoperto che spesso è possibile effettuare una diagnosi semplicemente guardando quali sono i test a cui è stato sottoposto

Dati mancanti (2)

- I possibili approcci quando si hanno dati con valori mancanti:
  - ignorare le istanze con valori mancanti
    - non molto efficace, in particolare se la percentuale di tuple con dati mancanti è alta.
    - si usa spesso quando il dato che manca è la classe in un problema di classificazione
  - riempire i valori mancanti manualmente
    - in generale è noioso, e potrebbe essere non fattibile
  - usare un valore costante come “Unknown” oppure 0 (a seconda del tipo di dati).
    - potrebbe alterare il funzionamento dell'algoritmo di analisi, meglio allora ricorrere ad algoritmi che gestiscono la possibilità di dati mancanti
    - è però utile se la mancanza di dati ha un significato particolare di cui tener conto

Dati mancanti (3)

- Altri possibili approcci:
  - usare la media dell'attributo al posto dei valori mancanti
  - per problemi di classificazione, usare la media dell'attributo per tutti i campioni della stessa classe
    - è una versione perfeziona del metodo della media per problemi di classificazione.
  - predirre il valore dell'attributo mancante sulla base degli altri attributi noti
    - la predizione può avvenire usando regressione lineare, alberi di classificazione, etc..
    - **si usano algoritmi di data mining per preparare i dati in input ad altri algoritmi di data mining.**

Dati inaccurati (1)

- Cause specifiche delle inesattezze
  - errori tipografici in attributi nominali: coca cola diventa coccola
  - inconsistenze: pepsi cola e pepsi
    - il sistema di data mining pensa si tratti di prodotti diversi
  - errori tipografici o di misura in attributi numerici
    - alcuni valori sono chiaramente poco sensati, e possono essere facilmente riconosciuti
    - ma altri errori possono essere più subdoli
  - errori deliberati: durante un sondaggio, l'intervistato può fornire un CAP falso
    - alcuni errori causati da sistemi di input automatizzati: se il sistema insiste per un codice ZIP (come il CAP ma negli USA) e l'utente non lo possiede?

Dati inaccurati (2)

- Occorre imparare a conoscere i propri dati!
  - capire il significato di tutti i campi
  - individuare gli errori che sono stati commessi
- Semplici programmi di visualizzazione grafica consentono di identificare rapidamente dei problemi:
  - la distribuzione è consistente con ciò che ci si aspetta?
  - c'è qualche dato ovviamente sbagliato?
- Vediamo due tecniche tipiche:
  - binning o regressione per eliminare il rumore
  - clustering: per riconoscere gli outliers

Dati rumorosi e binning

- Per rumore si intende un errore causale su una variabile misurata (tipicamente numerica)
  - è una delle possibile cause di dati inaccurati
- Il rumore può essere dovuto a
  - apparati di misura difettosi
  - problemi con le procedure di ingresso dati
  - problemi di trasmissione
  - limitazioni tecnologiche
- Il binning è una tecnica per ridurre la variabilità (e quindi il rumore) nei dati

Equi-Depth Binning (1)

- si considerano tutti i possibili valori (con ripetizioni) assunti dall'attributo e li si ordina
  - chiamiamo a_i con $i \in [1..N]$ i dati input, già ordinati
- si fissa un valore d per la profondità (depth) e si divide l'intervallo $[a_0, a_N]$ in intervalli (bin) consecutivi disgiunti, ognuno dei quali contenente all'incirca d elementi
  - quindi ci saranno circa N/d intervalli
  - chiamiamoli $I_1, ..., I_m$
  - la corrispondenza tra i dati e gli intervalli è data da una funzione v tale che $a_i \in I_{v(i)}$
- ora sostituiamo ad ogni a_i un valore derivato dal corrispondente intervallo

Equi-Depth Binning (2)

- varie possibilità per questa sostituzione
  - smoothing by bin means
    - si sostituisce ad a_i la media del corrispondente intervallo
    - $a_i \rightarrow$ media $I_{v(i)}$
  - smoothing by bin medians
    - si sostituisce ad a_i la mediana del corrispondente intervallo
    - $a_i \rightarrow$ mediana $I_{v(i)}$
  - smoothing by bin boundaries
    - si sostituisce ad a_i uno dei due estremi dell'intervallo corrispondente, in particolare quello più vicino
    - se $a_i - \min I_{v(i)} < \max I_{v(i)} - a_i$,
      allora $a_i \rightarrow \min I_{v(i)}$
      altrimenti $a_i \rightarrow \max I_{v(i)}$

Equi-Depth Binning (3)

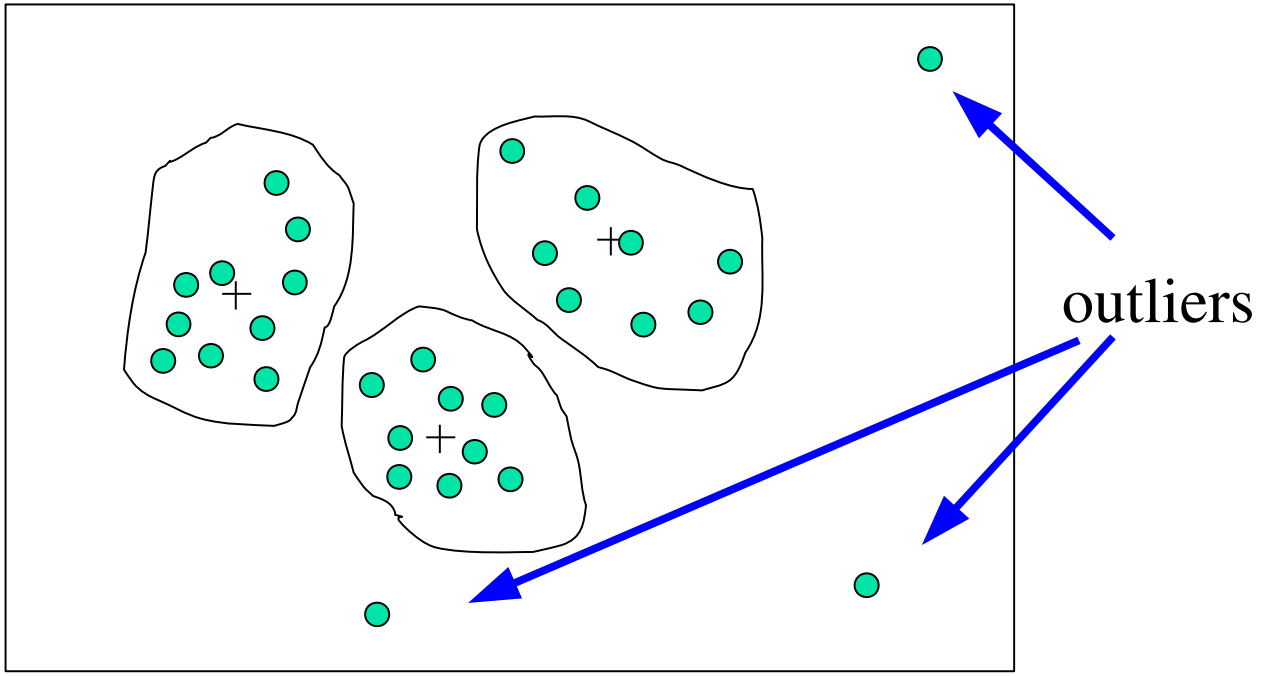
- Prezzi (in euro): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
  - Partizionamento in intervalli di 4 elementi (d=4)
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
  - Smoothing by bin means:
    - Bin 1: 9, 9, 9, 9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29
  - Smoothing by bin boundaries
    - Bin 1: 4, 4, 4, 15
    - Bin 2: 21, 21, 25, 25
    - Bin 3: 26, 26, 26, 34
- Non sempre è possibile avere intervalli di esattamente d elementi.

Equi-Width Binning e Regressione

- È simile all'Equi-Depth Binning, ma gli intervalli sono ottenuti in modo da avere più o meno tutti la stessa ampiezza (width)
  - ovvero, se gli intervalli che otteniamo sono I_1, ..., I_m, il valore $\max I_i - \min I_i$ è più o meno costante, al variare di $i \in [1...m]$
- Con i dati di prima, e una ampiezza per ogni intervallo più o meno fissata a 10, otteniamo i seguenti bin:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25, 26, 28, 29, 34
- Un altro metodo per ammorbidire i dati è determinare una funzione che li approssimi, e sostituire i valori previsti dalla funzione a quelli effettivi
  - ad esempio si può usare la regressione lineare

Clustering

- Con questo metodo è possibile riconoscere gli outliers.
  - si dividono i possibili valori degli attributi da pulire in gruppi;
  - eventuali valori che non ricadono in nessun gruppo sono degli outliers.
- Anche in questo caso usiamo algoritmi di data-mining come preparazione per altri algoritmi di data-mining



Data Integration

Data Integration (1)

- Si tratta di combinare dati provenienti da sorgenti diverse
- Primo aspetto: schema integration
  - integrazione degli schemi relazionali tra database diversi
  - il problema è capire che relazione c'è tra entità provenienti da diverse sorgenti
    - ad esempio, come fa il progettista a capire se l'attributo customer_id di un database e cust_number in un altro si riferiscono alla stessa entità
- Secondo aspetto: integrazione dei dati vera propria
  - ammesso di aver trovato che le tabelle customer di due diversi database si riferiscono alla stessa entità, come si fa a mettere assieme in una unica tabella le informazioni?
  - possibilità di informazioni discordanti (errori, unità di misura diverse, etc..)

Data Integration (2)

- Terzo aspetto: ridondanza
  - alcuni attributi possono essere ricavati (perfettamente o in parte) da altri
    - ad esempio, non si capisce che i campi “categoria merceleogica” di una tabella è “tipo prodotto” di un altra si riferiscono allo stesso tipo di informazione.
    - quando si metteno assieme i dati, si creano due campi diversi nella tabella integrata.
    - i valori dei due campi sono strettamente collegati
  - per gli attributi numerici, è possibile provare a scoprire se due attributi sono tra loro ridondanti usando una analisi di correlazione (che vedremo in un'altra lezione)
  - in aggiunta alla ridondanza tra attributi, va anche controllata la ridondanza tra tuple, evitando di creare delle tuple duplicate

Data Trasformation

Data Transformation

- I dati sono consolidati e trasformati in forme più appropriate per le analisi. Varie possibilità sono
  - smoothing (lisciamento) rimuovere i rumori nei dati (binning, clustering, regressione, ...)
    - già visti nella fase di Data Cleaning
  - aggregazione: costruire dati aggregati prima dell'analisi
    - nei sistemi OLAM, corrisponde a scegliere il cuboide appropriato
    - eventualmente usando le gerarchie di concetti
  - costruzione degli attributi: costruire nuovi attributi a partire da quelli presenti per aiutare l'algoritmo di analisi
    - per esempio, si aggiunge l'attributo derivato area come prodotto degli attributi altezza e larghezza.
  - normalizzazione: modificare la scala dei dati in modo che cadano in intervalli stabiliti (ad esempio da -1ad 1)

Normalizzazione (1)

- Spesso gli attributi assumono valori in intervalli di ampiezza diversa.
  - può compromettere il funzionamento di alcune analisi
  - necessità di normalizzare i dati
- min-max normalization: si riscala l'attributo A in modo che i nuovi valori cadano tra new_min_A e new_max_A.

$$v' = \frac{v - \boldsymbol{min}_A}{\boldsymbol{max}_A - \boldsymbol{min}_A}\left(\boldsymbol{new}_{\boldsymbol{max}_A} - \boldsymbol{new}_{\boldsymbol{min}_A}\right) + \boldsymbol{new}_{\boldsymbol{min}_A}$$

  - il minimo e il massimo effettivo dell'attributo A potrebbero essere ignoti.
    - si verifica un superamento dei nuovi limiti se successivamente appare un dato con un valore di A oltre l'intervallo originario.
  - molto influenzato dagli outliers

Normalizzazione (2)

- z-score normalization (anche z-mean normalization)

$$v' = \frac{v - mean_A}{\sigma_A}$$

media

deviazione standard

  - utile quando non si conosce minimo e massimo per A
  - i valori normalizzato non hanno un minimo e un massimo fissato
  - non influenzato dagli outlier (o almeno non altrettanto del metodo precedente)

Normalizzazione (3)

- normalizzazione per scalatura decimale:
  - una variante del metodo min-max : restringe i valori tra -1 ed 1 modificando la posizione della virgola

$$v' = \frac{v}{10^j}$$ dove j è il più piccolo intero tale che Max(|v' |)<1

  - ad esempio, se l'attributo A varia da -986 a 917, per normalizzare dividiamo tutto per 1000. I nuovi valori andranno da -0.986 a 0.917.
  - il passaggio da valori di base a valori normalizzati è molto semplice

Data Reduction

Data Reduction

- I data warehouse possono memorizzare dati la cui dimensione è dell'ordine dei terabyte: le analisi sono troppo complesse.
- Necessità di effettuare una riduzione dei dati
  - ottenere una rappresentazione ridotta dei dati con una occupazione molto inferiore di memoria ma che produce gli stessi (o comunque simili) risultati analitici.
- Varie strategie
  - aggregazione
    - usare un cuboide a più alto livello di aggregazione, purché sufficiente per il compito di analisi che dobbiamo svolgere.
  - riduzione della dimensionalità (dimensionality reduction)
    - selezione di attributi rilevanti
  - riduzione della numerosità (numerosity reduction)
  - discretizzazione e generazione delle gerarchie di concetto

Selezione di attributi rilevanti (1)

- Selezionare un insieme minimo di attributi che descrivano in maniera adeguata i dati in ingresso
  - ad esempio, eliminare gli attributi irrilevanti, come può essere una chiave primaria
- Può essere effettuata da un esperto del settore sotto analisi, ma qui parleremo di metodi automatici
- Serve una "misura" della bontà di un insieme di attributi, in modo che si possa scegliere l'insieme migliore.
- Una ricerca esaustiva è spesso impossibile:
  - se ho d attributi in totale, ci sono 2^d possibili sottoinsiemi
  - si usano quindi algoritmi euristici

Selezione di attributi rilevanti (2)

- Possibili algoritmi euristici
  - step-wise forward selection
    - parto da un insieme vuoto di attributi
    - ad ogni passo aggiungo l'attributo che massimizza la qualità dell'insieme risultante
      - Insieme di attributi : {A1, A2, A3, A4, A5, A6 }
      - Insiemi ridotti: {} → {A1} → {A1,A6}→ {A1,A4,A6}
  - step-wise backward selection
    - parto da tutti gli attributi
    - ad ogni passo tolgo l'attributo che massimizza la qualità dell'insieme risultante
      - Insieme di attributi : {A1, A2, A3, A4, A5, A6 }
      - Insiemi ridotti: {A1, A2, A3, A4, A5, A6} → {A1, A3, A4, A5, A6} → {A1, A4, A5, A6} → {A1, A4, A6}
  - combinazione di forward e backward selection

Selezione di attributi rilevanti (3)

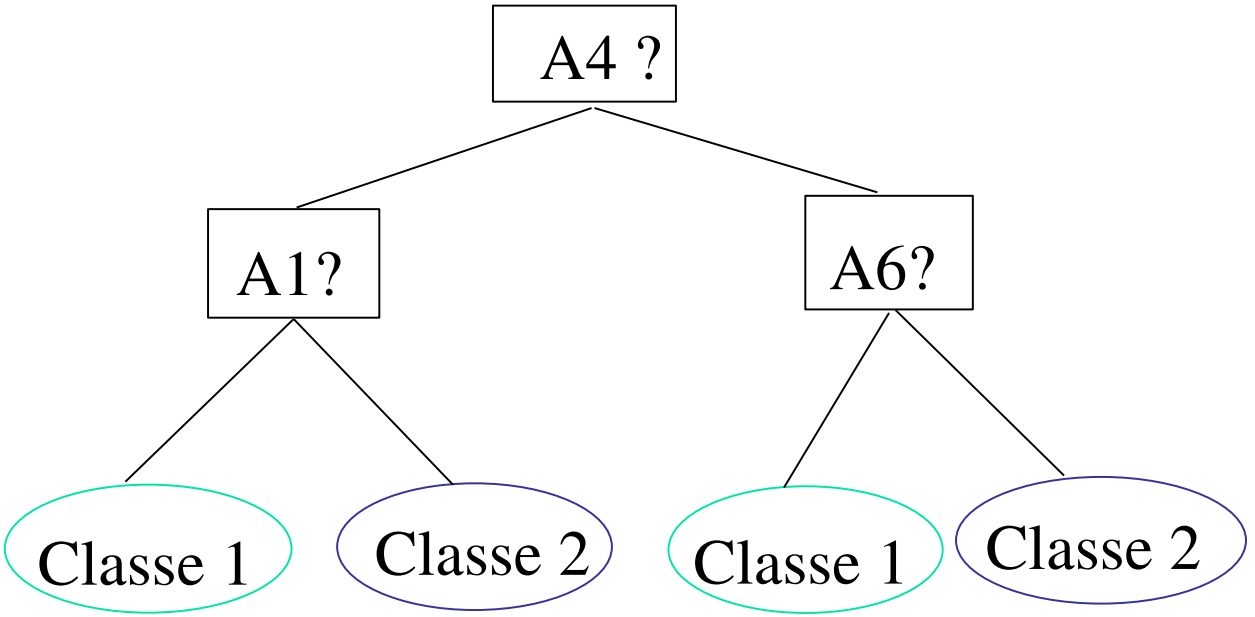
- Necessità di stabilire dei criteri di arresto. Ci si ferma quando:
  - si è raggiunto un numero fissato di attributi
  - forward selection
    - si è raggiunta una qualità minima desiderata per l'insieme degli attributi
    - l'incremento minimo di qualità scende sotto una determinata soglia
  - backward selection:
    - si è scesi sotto la qualità minima desiderata per l'insieme degli attributi
    - c'è una riduzione brusca della qualità
- Cosa usare come misura di qualità di un sottoinsieme di attributi?
  - una possibilità è usare il guadagno di informazione, eventualmente corretto per penalizzare gli insiemi con attributi strettamente correlati.

Selezione di attributi rilevanti (4)

- In alternativa a questi metodi euristici, posso usare gli alberi di decisione: costruisco un alberi di decisione e mantengo solo gli attributi che in esso appaiono.

Insieme iniziale di attributi:

{A1, A2, A3, A4, A5, A6}



Insieme ridotto: {A1, A4, A6}

Compressione dei dati

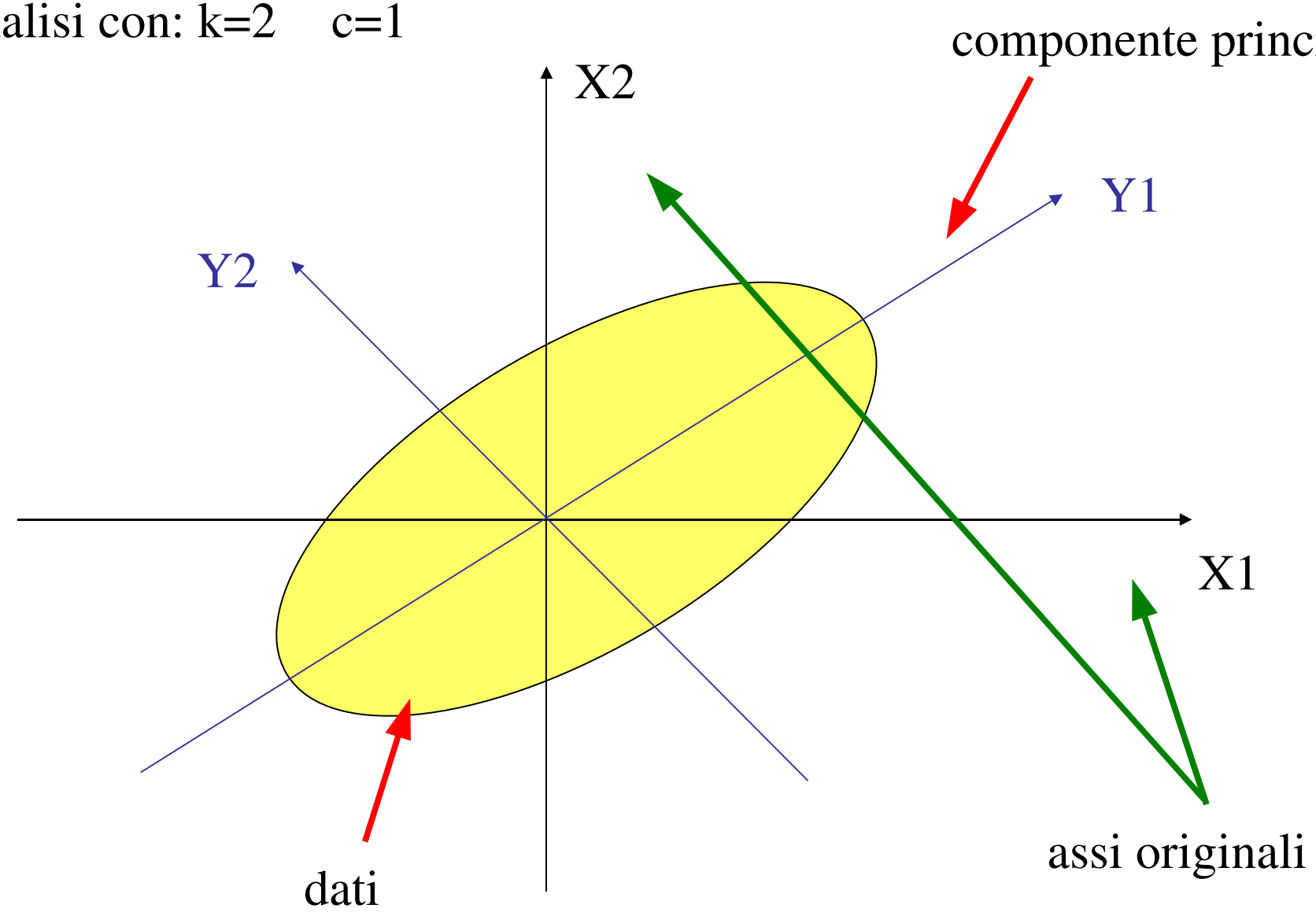
- Un insieme di tecniche che consentono di ridurre la dimensioni dei dati
  - lossless (senza perdita)
    - una enorme letteratura per quello che riguarda la compressione delle stringhe, ma i dati ottenuti non possono essere manipolati direttamente
    - esempio: formati ZIP, RAR, Gzip, BZ2 per file generici
    - esempio: formati GIF per immagini
  - lossy (con perdita)
    - tipicamente usate per contenuti multimediali
    - esempio: mp3, Ogg Vorbis (audio), MPEG (video), JPEG (immagini)
- Tecniche più usate nelle applicazioni di analisi dei dati:
  - DFT / DWT (Discrete Fourier/Wavelet transform)
  - PCA (Principal Component Analysis)

PCA (1)

- Analisi delle componenti principali
  - dati N vettori in k dimensioni, trovare c<=k vettori ortonormali (le componenti principali) che possono essere usati per rappresentare i dati
    - l'insieme di dati originale viene ridotto ad un insieme di N vettori in c dimensioni
  - i nuovi dati sono combinazioni lineari delle c componenti principali
    - le componenti principali sono ordinate per "significatività"
    - le più significative sono quelle che mostrano maggiore variabilità nei dati
    - le meno significativi sono quelli che mostrano minore variabilità nei dati
  - si può decidere di eliminare le componenti meno significative
  - funziona solo su dati numerici, che devono essere prima normalizzati.

PCA (2)

analisi con: k=2 c=1



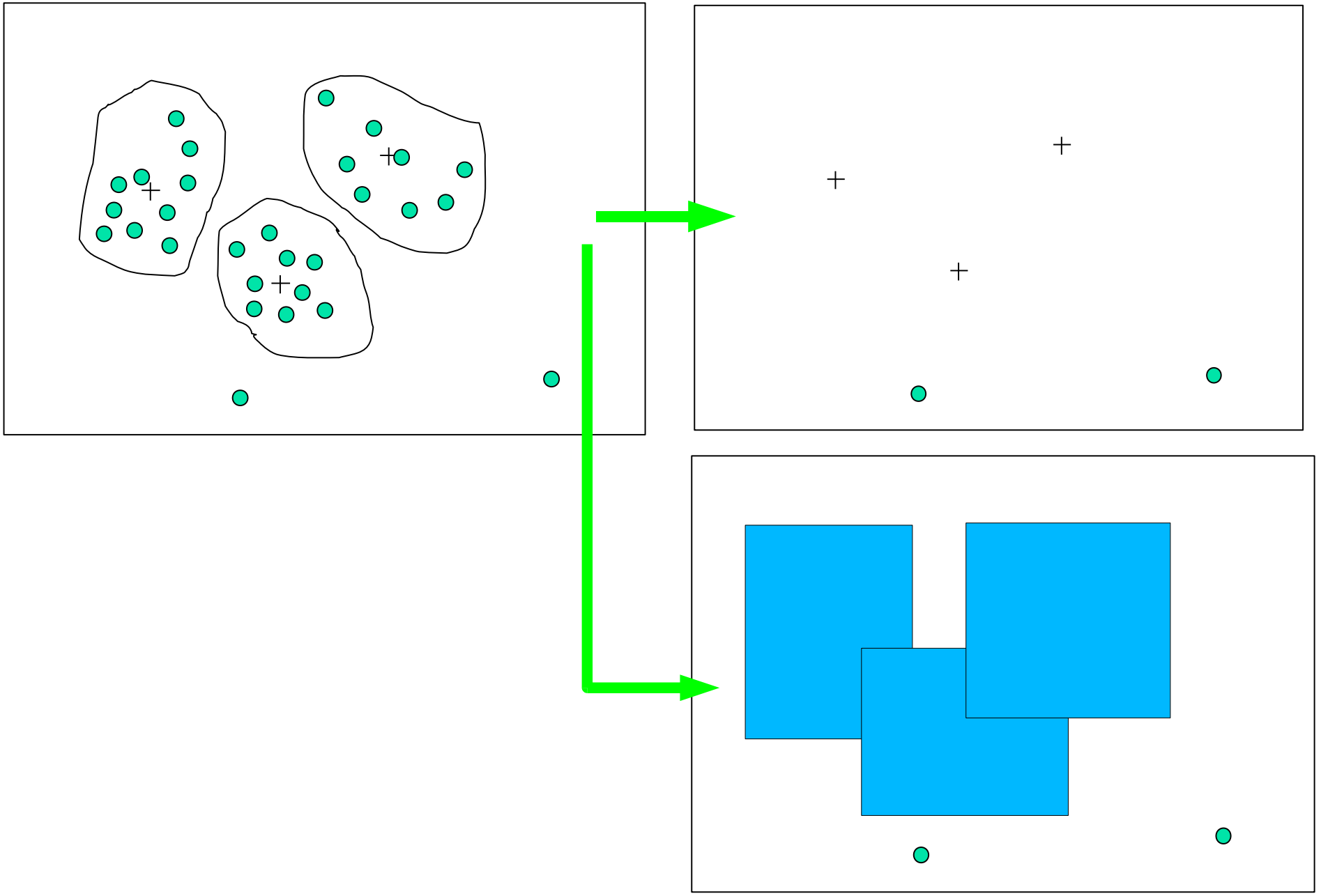
Riduzione della numerosità

- Ridurre la mole dei dati scegliendo una rappresentazione diversa
- Metodi parametrici
  - assumere che i dati soddisfino un dato modello, stimare i parametri del modello e usare questi ultimi invece dei parametri originali
    - ad esempio, stimare una serie di numeri usando la regressione lineare..
    - od un insieme di numeri con una distribuzione gaussiana
  - assumiamo che siano ben noti dai corsi di statistica
- Metodi non parametrici
  - non si assume nessun modello particolare
  - famiglie principali di metodi:
    - raggruppamento
    - campionamento

Raggruppamento (1)

- Si dividono i dati in cluster
- La rappresentazione dei cluster sostituisce la rappresentazione iniziale dei dati.
  - cosa si intende per rappresentazione dei cluster?
  - varie possibilità:
    - un punto medio,
    - una figura geometrica che approssima il cluster (poligoni, cerchi, etc..)

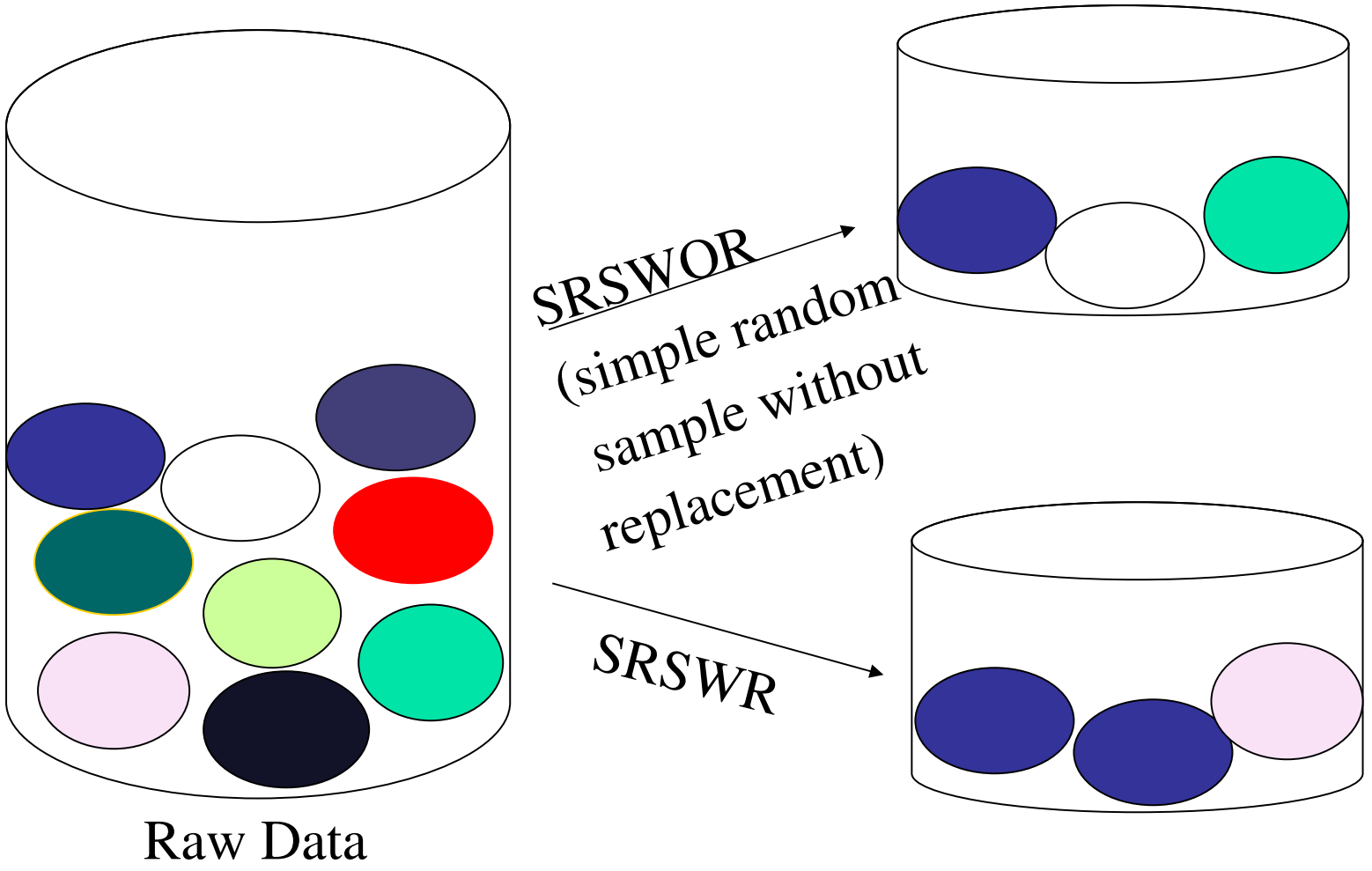
Raggruppamento (2)



Campionamento (1)

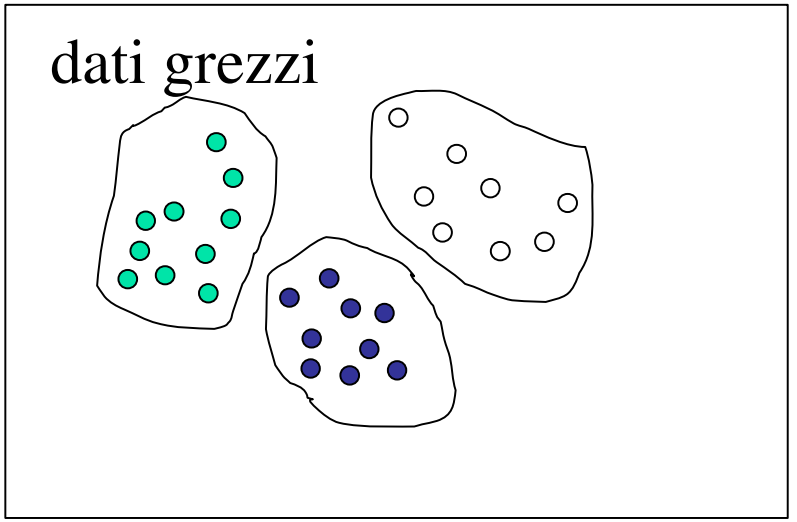
- Scegliere un sottoinsieme dei dati (campione) per eseguire le analisi.
- Sia N il numero di istanze nel mio insieme di dati D e n il numero di istanze che voglio scegliere per il campione.
  - campionamento semplice senza rimpiazzo:
    - scelgo n < N istanze da D. In ogni scelta, tutte le istanze hanno uguale probabilità, e non è possibile scegliere due volte la stessa istanza.
  - campionamento semplice con rimpiazzo
    - come sopra ma è possibile scegliere più volte la stessa istanza
- Un vantaggio dei metodi di campionamento è che il costo per ottenere un campione è proporzionale alla dimensione del campione
  - ha dunque complessità sub-lineare rispetto ai dati in input

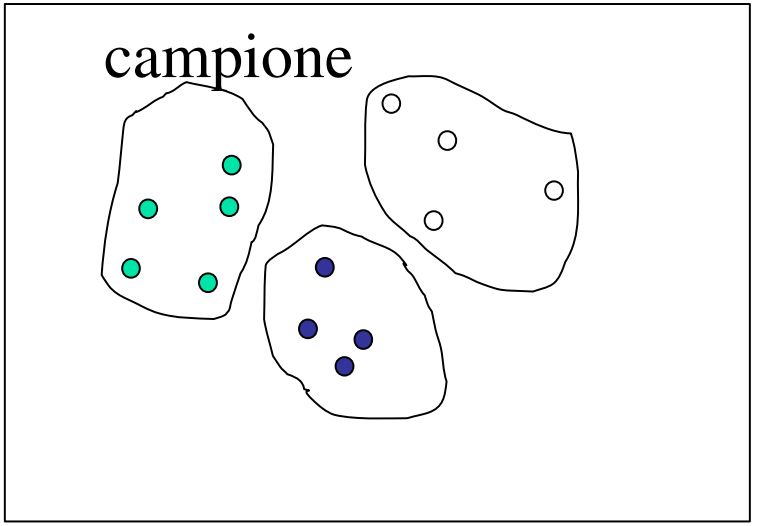
Campionamento (2)



Campionamento (3)

- Campionamento stratificato
  - i dati sono divisi in gruppi disgiunti $G_1 \ldots G_k$ chiamati strati
  - scelgo da ogni gruppo G_i un campione proporzionale alla dimensione di G_i (ovvero scelgo $n*|G_i|/N$ elementi)
  - per problemi di classificazione, gli stati corrispondono alle classi
  - utile quando alcune classi sono rappresentate da un numero esiguo di istanze, per assicurarci che nel campione sia presente qualcuna di esse.





Discretizzazione e gerarchie di concetti

Discretizzazione

- Consiste nel ridurre il numero di possibili valori diversi che assume un attributo
  - si divide il range dell'attributo in intervalli o comunque in sottoinsiemi
  - si sostituisce ai dati originali una etichetta che rappresenti l'intervallo o il sottoinsieme a cui appartiene
  - può migliorare l'efficienza di alcuni algoritmi di analisi
- Si può applicare la discretizzazione in maniera ricorsiva per ottenere una gerarchia di concetti
- Alcuni metodi per valori numerici
  - binning (visto già come tecnica di data cleaning) e istogrammi
  - analisi di raggruppamento
  - discretizzazione basata sull'entropia
  - segmentazione per partizionamento naturale

Istogrammi

- Si procede come per il binning
  - invece di rimpiazzare i dati con un valore caratteristico del bin a cui esso appartiene (media, mediana, estemi o altro), lo si sostituisce con una "etichetta" che rappresenta l'intervallo di valori che esso assume
  - Esempio: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
    - Partizionamento equi-depth in intervalli di 4 elementi (d=4)
      - Bin 1: 4, 8, 9, 15
      - Bin 2: 21, 21, 24, 25
      - Bin 3: 26, 28, 29, 34
    - Intervalli ottenuti: [-4,18),[18,25.5),[25.5,34]
    - gli estremi di ogni intervallo sono stati posti a metà tra il massimo valore del bin corrispondente e il minimo del successivo.

Discretizzazione ed entropia (1)

- Si applica tipicamente come preliminare alla classificazione
  - esiste un attributo “classe” C per il calcolo dell'entropia
- Ogni valore *v* di un attributo A è una possibile frontiera per la divisione negli intervalli $A \leq v$ e $A > v$.
- Scelgo il valore che mi da il maggiore guadagno di informazione IG(A,*v*)
- Il processo si applica ricorsivamente ai sotto-intervalli così ottenuti, fino a che non si raggiunge una condizione di arresto
  - ad esempio, fino a che il guadagno di informazione che si ottiene diventa inferiore a una certa soglia *d*

Discretizzazione ed entropia (2)

- Supponiamo di avere le classi “s” ed “n” e l'insieme S costituito dagli attributi A e C: (0, s), (2,n), (30,n), (31,n), (32,s), (40,s).
  - Calcoliamo i diversi guadagni di informazione
    - IG(A,0)=0.19 E(S,2)=0 E(S,30)=0.08 E(S,31)=0.46
    - E(S,32)=0.19 E(S,40)=0
  - Il guadagno di informazione maggiore si ha generando i sotto-intervalli $A \leq 30$ e $A > 30$
- Spesso, piuttosto che dividere sul valore ottenuto (30), si divide sul valore di mezzo tra quello ottenuto e il successivo
  - in questo caso si avrebbe $A \leq 31.5$ e $A > 31$

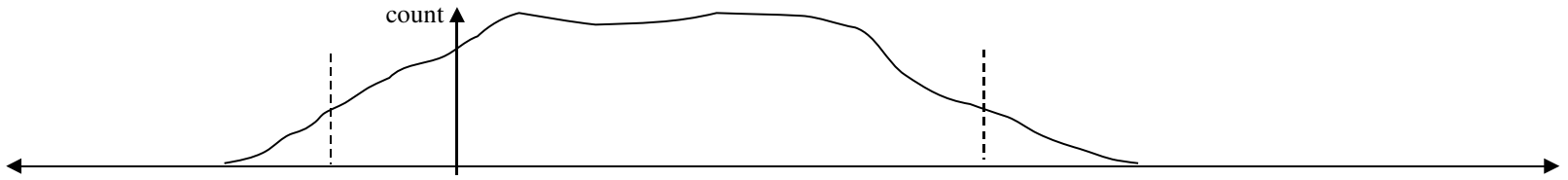
Partizionamento naturale (1)

- I metodi precedenti danno spesso intervalli innaturali
  - intervalli del tipo (€50.000-€60.000) sono più desiderabili di intervalli del tipo (€51.492-€60.872).
- la regola 3-4-5 può essere usata per generare intervalli naturali
- in linea di massima, il metodo divide un dato intervallo in 3, 4 o 5 sotto-intervalli diversi, ricorsivamente a seconda dell'intervallo di valori assunti dalla cifra più significativa
  - se l'intervallo copre 3, 6, 7 o 9 valori distinti della cifra più significativa, viene partizionato in 3 sotto-intervalli
    - di uguale ampiezza per 3, 6 o 9 valori di stinti
    - nella proporzione 2-3-2 per 7 valori distinti
  - per 2, 4 o 8 valori distinti, partiziona i dati in 4 intervalli.
  - per 1, 5 o 10 valori distinti, partiziona i dati in 5 intervalli.

Partizionamento naturale (2)

- Se l'attributo A varia da -199 a 1838:
  - si arrotondano gli estremi dell'intervallo alla cifra più significativa, ottenendo l'intervallo (-1000,2000]
  - l'intervallo copre 3 cifre significative diverse
    - (2000-(-1000))/1000=3
    - si divide in 3 sotto-intervalli (-1000,0], (0,1000], (1000,2000]
  - si procede ricorsivamente (se si vuole)
- la regola funziona male se ci sono valori estremi molto diversi dai valori medi.. in tal caso, si possono usare per partizionare solo i dati dal 5° al 95° percentile

Partizionamento naturale (3)

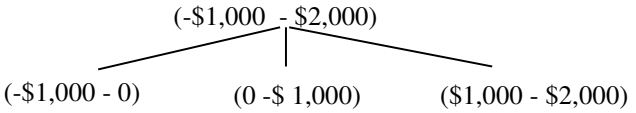


Step 1: -$351 -$159 profit $1,838 $4,700

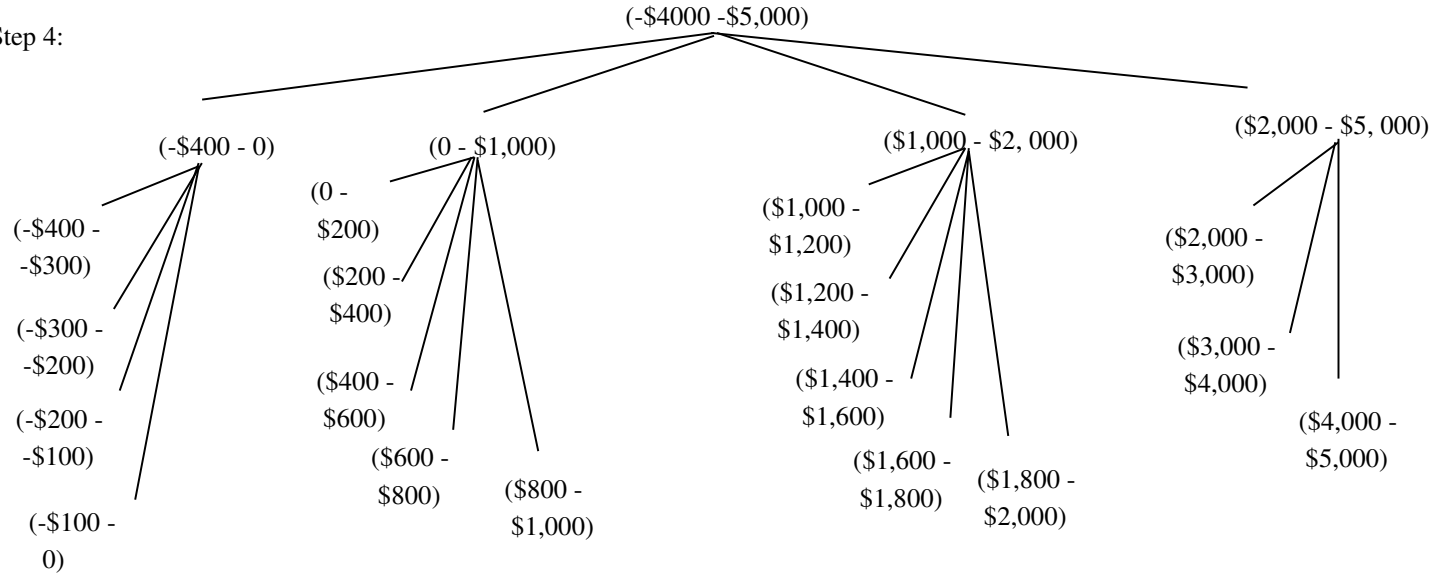
Min Low (i.e, 5%-tile) High(i.e, 95%-0 tile) Max

Step 2: msd=1,000 Low=-$1,000 High=$2,000

Step 3:



Step 4:



Gerarchie di concetti per attributi categoriali (1)

- I metodi visti prima funzionano per attributi numerici
  - e per gli attributi categoriali?
- Abbiamo già visto i concetti di
  - schema hierarchy
  - set-grouping hierarchy
- Esistono possibilità intermedie
  - specificare gli attributi da usare per la gerarchia ma non l'ordinamento.
    - l'ordinamento è ricercato automaticamente, basandosi sul numero dei distinti valori assunti dagli attributi
  - specificare parte degli attributi da usare per la gerarchia
    - gli altri sono automaticamente inseriti dal sistema, sfruttando eventuale informazione semantica dello schema del database

Gerarchie di concetti per attributi categoriali (2)

- Esempio
  - se l'utente specifica gli attributi country, state, city e street, il sistema ne trova la relazione analizzando i valori distinti;
  - se l'utente specifica solo city, gli altri vengono inseriti automaticamente perchè considerati legati dal punto di vista semantico.