Data Mining

Tecniche e algoritmi di base per l'estrazione di conoscenza

Cosa vedremo ...

- Motivazione al data mining
- Panoramica approcci di base
- Idee generali algoritmi di base
- Collegamenti con discipline coinvolte
- Esempi applicativi

Cosa non vedremo ...

- Algoritmi nello specifico
- Tecniche avanzate

Data mining

Introduzione

Knowledge Discovery

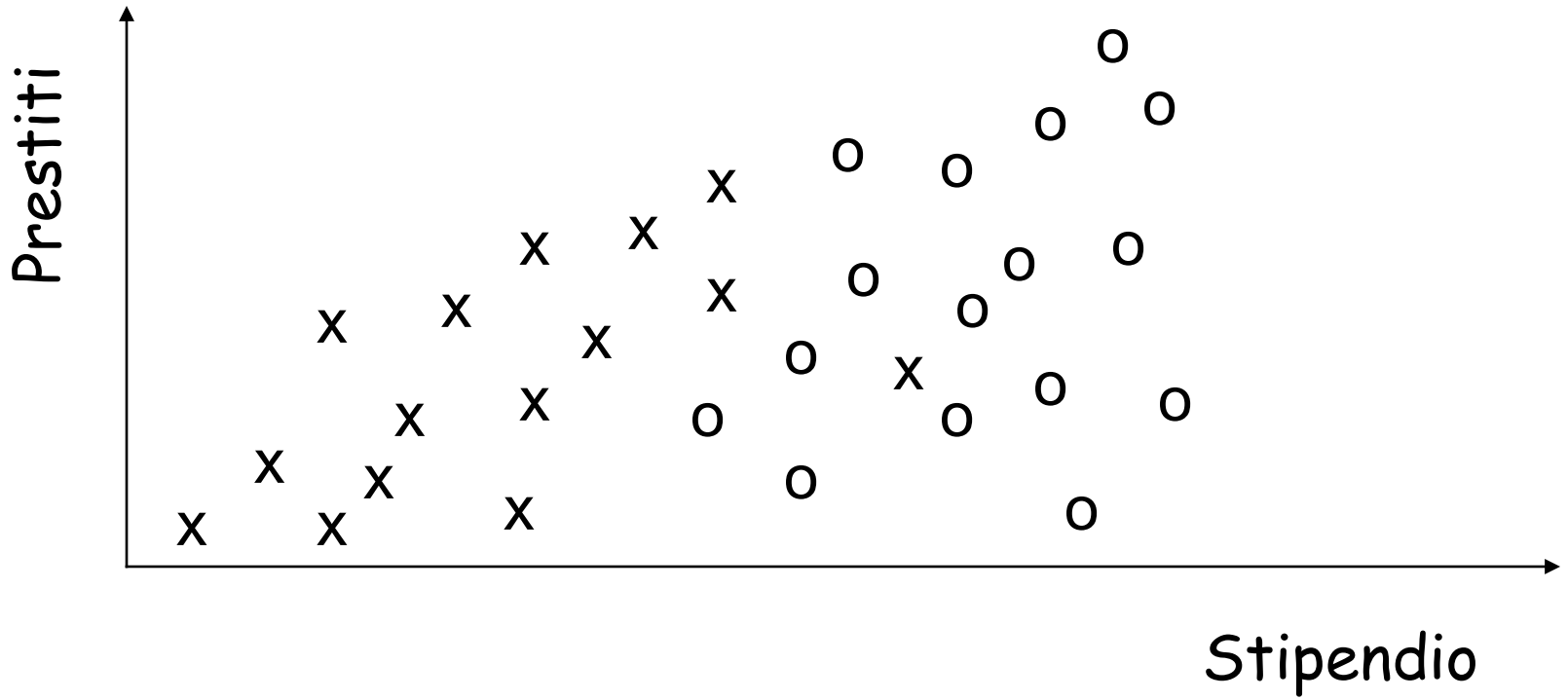
- La maggior parte delle aziende dispone di enormi basi di dati contenenti dati di tipo operativo
  - Queste basi di dati costituiscono una potenziale miniera di utili informazioni

Knowledge Discovery

- Processo di estrazione dai dati esistenti di pattern:
  - valide
  - precedentemente sconosciute
  - potenzialmente utili
  - comprensibili

    [Fayyad, Piatesky-Shapiro, Smith 1996]

Esempio



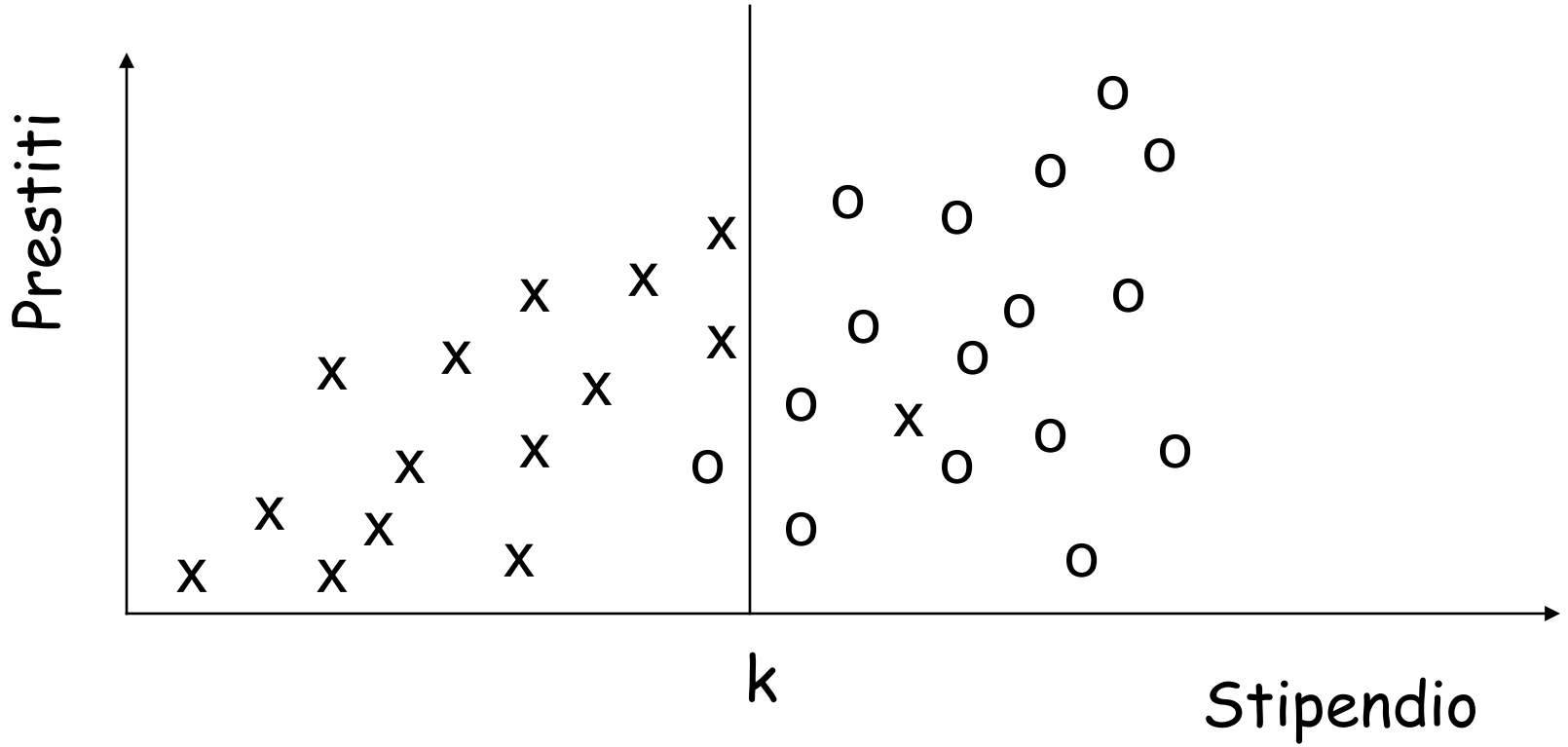
Persone che hanno ricevuto un prestito dalla banca:
x: persone che hanno mancato la restituzione di rate
o: persone che hanno rispettato le scadenze

Knowledge Discovery

- *Dati*: insieme di informazioni contenute in una base di dati o data warehouse

- *Pattern*: espressione in un linguaggio opportuno che descrive in modo succinto le informazioni estratte dai dati
  - regolarita`
  - informazione di alto livello

Esempio



IF stipendio < k THEN mancati pagamenti

Caratteristiche dei pattern

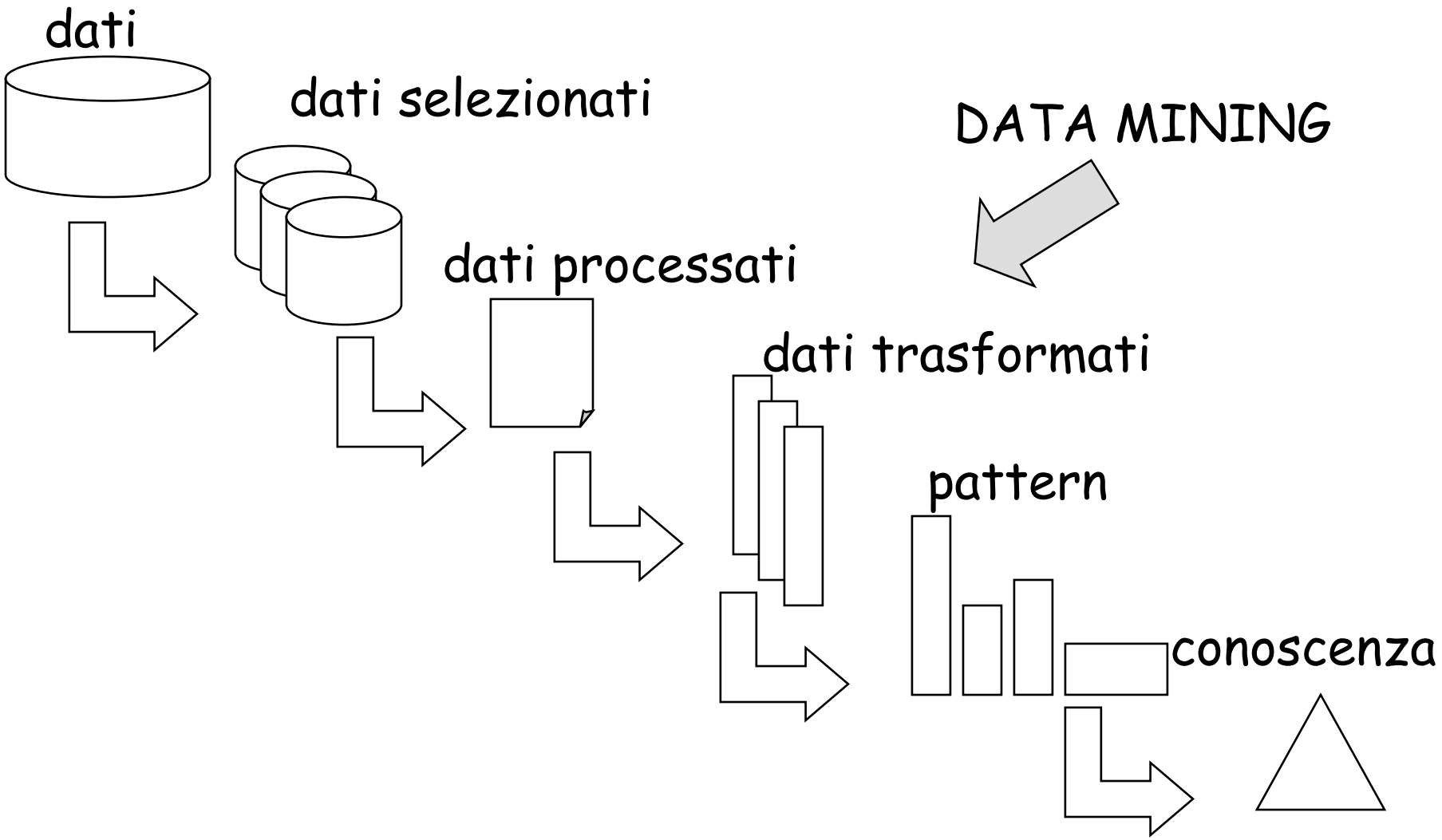
- *Validita`*: i pattern scoperti devono essere validi su nuovi dati con un certo grado di certezza
  - Esempio: spostamento a destra del valore di k porta riduzione del grado di certezza

- *Novita`*: misurata rispetto a variazioni dei dati o della conoscenza estratta

Caratteristiche dei pattern

- *Utilita`*
  - Esempio: aumento di profitto atteso dalla banca associato alla regola estratta
- *Comprensibilita`*: misure di tipo
  - sintattico
  - semantico

Processo di estrazione



Requisiti

- Dati eterogenei
- efficienza e scalabilita` tecniche
- qualita` informazioni estratte
- criteri diversificati di estrazione
- privatezza dei dati

Discipline coinvolte

- AI
- machine learning
- knowledge acquisition
- statistics
- data visualization
- neural networks
- database
- data mining

Dati in memoria centrale

Dati in memoria secondaria

Data Mining

Tecniche di analisi

Tecniche di analisi

- Regole di associazione
- Classificazione
- Clustering
- Similarity search

Regole di associazione

- Dati del problema:
  - I insieme di items
    - prodotti venduti da un supermercato
  - *transazione T*: insieme di items t.c. $T \subseteq I$
    - oggetti acquistati nella stessa transazione di cassa al supermercato
  - *base di dati D*: insieme di transazioni

Regole di associazione

- Regola di associazione $X \Rightarrow Y$

  $X, Y \subseteq I$

- *Supporto S*: $\frac{\#\text{trans. contenenti } X \cup Y}{\#\text{trans. in D}}$
  - rilevanza statistica
- *Confidenza C:* $\frac{\#\text{trans. contenenti } X \cup Y}{\#\text{trans. contenenti } X}$
  - significativita` dell'implicazione

Esempio

Latte $\Rightarrow$ Uova

- Supporto: il 30% delle transazioni che contengono latte contiene anche uova
- Confidenza: il 2% delle transazioni contiene entrambi gli elementi

Regole di associazione

- Problema:
  - determinare tutte le regole con supporto e confidenza superiori ad una soglia data

Esempio

TRANSACTION ID	OGGETTI ACQUISTATI
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

- Assumiano:
  - supporto minimo 50%
  - confidenza minima 50%

Esempio

TRANSACTION ID	OGGETTI ACQUISTATI
2000	**A**,B,**C**
1000	**A**,**C**
4000	**A**,D
5000	B,E,F

- Regole ottenute:
  - *A* ⇒ *C* supporto 50% confidenza 66.6
  - *C* ⇒ *A* supporto 50% confidenza 100%

Applicazioni

- Analisi market basket
  - * ⇒ uova
    - cosa si deve promuovere per aumentare le vendite di uova?
  - Latte ⇒ *
    - quali altri prodotti devono essere venduti da un supermercato che vende latte?
- Dimensione del problema:
  - oggetti: 10^4, 10^5, transazioni: > 10^6
  - base di dati: 10-100 GB

Settori di sviluppo

- Teniche definite prevalentemente nel contesto data mining

Decomposizione problema

- Trovare tutti gli insiemi di item che hanno un supporto minimo (*frequent itemsets*)
- Generazione delle regole a partire dai frequent itemsets
- Algoritmo fondamentale: APRIORI
  [Agrawal, Srikant 1994]

Esempio

- Passo 1: estrazione frequent itemsets

TRANSACTION ID	OGGETTI ACQUISTATI
1	A,B,C
2	A,C
3	A,D
4	B,E,F

- supporto minimo 50%

FREQUENT ITEMSET	SUPPORTO
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

Esempio

- Passo 2: estrazione regole
  - confidenza minima 50%
  - Esempio: regola $A \Rightarrow C$
    - supporto $\{A,C\}$ = 50%
    - confidenza = supporto $\{A,C\}$/supporto$\{A\}$ = 66.6%
  - regole estratte
    - $A \Rightarrow C$ supporto 50%, conf. 66.6%
    - $A \Rightarrow C$ supporto 50%, conf. 100%

Interesse regole estratte

- Non sempre tutte le regole con supporto e confidenza superiori alla soglia sono interessanti

- Esempio:
  - scuola con 5000 studenti
  - 60% (3000) gioca a pallacanestro
  - 75% (3750) mangia fiocchi a colazione
  - 40% (2000) gioca a pallacanestro e mangia fiocchi a colazione

Interesse regole estratte

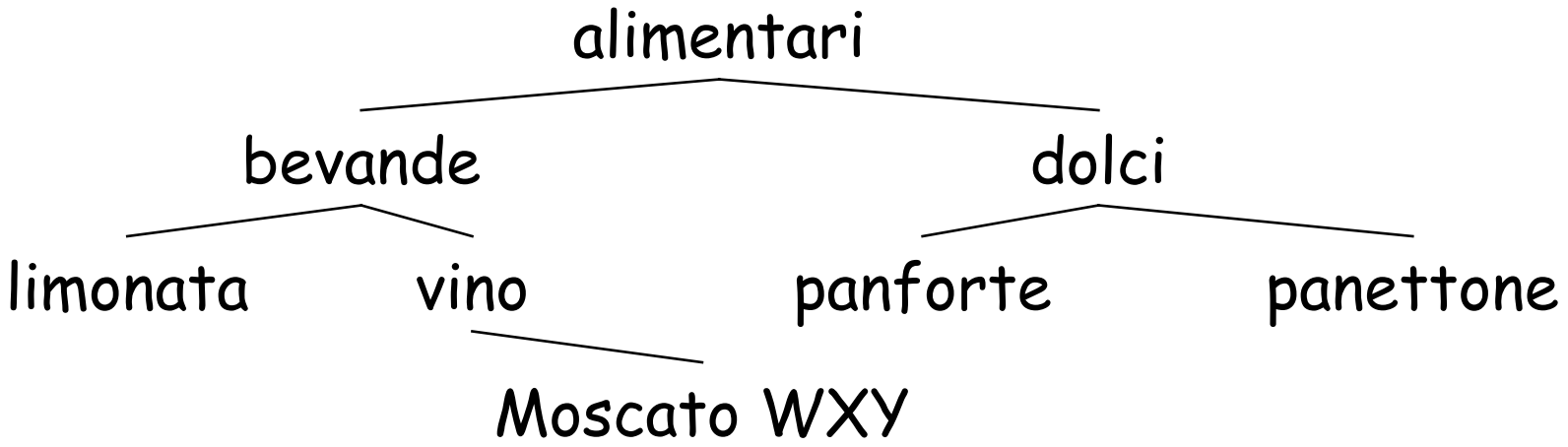
- Con supporto min. 40% e confidenza min. 60%:
  - gioca a pallacanestro ⇒ mangia fiocchi
  - supporto = 2000/5000 = 0.4
  - confidenza = 2000/3000 = 0.66 > 0.6

- regola fuorviante perche` il 75% degli studenti mangia fiocchi!

- Nuova misura: $\frac{\text{supporto}(A,B) - \text{supporto}(B)}{\text{supporto}(A)}$

Estensioni

- Regole generalizzate



- vino ⇒ dolci, valida anche se:
  - bevande ⇒ dolci non ha confidenza suff.
  - Moscato WXY ⇒ dolci non ha supporto suff.

Estensioni

- Estrazione da dati di tipo numerico
  - partizionamento in intervalli
  - misura della perdita di informazione dovuta al partizionamento
    - sposato, eta` 50-60 anni ⇒ 2 automobili
- Metaregole
- Ottimizzazione algoritmi:
  - riduzione scan database
  - campionamento
  - update incrementale
  - algoritmi paralleli

Classificazione

- Dati del problema:
  - insieme di classi
  - insieme di oggetti etichettati con il nome della classe di appartenenza (*training set*)

- Problema:
  - trovare il profilo descrittivo per ogni classe, utilizzando le features dei dati contenuti nel training set, che permetta di assegnare altri oggetti, contenuti in un certo *test set*, alla classe appropriata

Applicazioni

- Classificazione tendenze di mercato
- identificazione automatica di immagini
- identificazione del rischio in mutui/assicurazioni
- efficiacia trattamenti medici

Settori di sviluppo

- Statistica
- machine learning
  - alberi di decisione
  - inductive logic programming
- reti neurali
- sistemi esperti
- data mining

Ipotesi di funzionamento

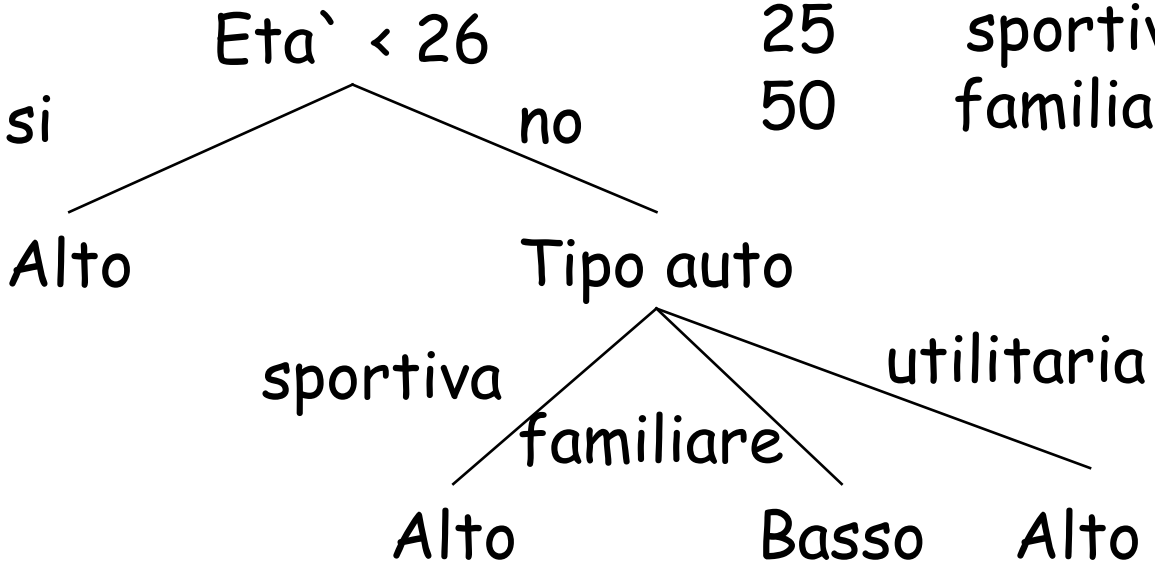
- Training set contenuto nella memoria principale del sistema
- Nelle DB attuali possono essere disponibili anche Mbyte di training set
  - dimensioni significative del training set possono migliorare l'accuratezza della classificazione

Alberi di decisione

- Veloci rispetto agli altri metodi
- Facili da interpretare tramite regole di classificazione
- Possono essere convertiti in interrogazioni SQL per interrogare la base di dati

Esempio

ETA`	TIPO AUTO	CLASSE RISCHIO
40	familiare	basso
65	sportiva	alto
20	utilitaria	alto
25	sportiva	alto
50	familiare	basso



Costruzione albero

- Due fasi:
  - *fase di build*: si costruisce l'albero iniziale, partizionando ripetutamente il training set sul valore di un attributo, fino a quando tutti gli esempi in ogni partizione appartengono ad una sola classe
  - *fase di pruning*: si pota l'albero, eliminando rami dovuti a rumore o fluttuazioni statistiche

Fase di build

Maketree(training set T)
{Partition(T)}

Partition(Data set S)
{
    if (tutti i punti in S sono nella stessa classe) then
    return
    for tutti gi attributi A do
    valuta gli splits su A
    usa split migliore per partizionare S in S1 e S2
    Partition(S1)
    Partition(S2)
}

}

Split

- Su attributi numerici
  - $A \leq v$
  - v1 v2 v3 v4 v5

    s1 s2 s3 s4 valori di split
- Su attributi categorici
  - $A = s1$, per ogni s1 nel dominio di A, se e` piccolo
  - euristica altrimenti

Indici di splitting

- Diverse tipologie di indice
- Indice Gini:
  - dataset T con esempi di n classi
  - $gini(T) = 1 - \Sigma_i p_i^2$
  - con p_i frequenza class i in T
- Se T viene suddiviso in T1 con n1 esempi e T2 con n2 esempi:
  - $gini_{split}(T) = (n1/n)\, gini(T1) + (n2/n)\, gini(T2)$

Fase di pruning

- Due modalita` di base
  - si estraggono campioni multipli dal training set e si costruiscono alberi indipendenti
    - gli alberi vengono utilizzati per stimare il tasso di errore dei sottolaberi dell'albero di partenza
    - non appropriato per training set molto grandi
  - si divide il training set in due parti:
    - costruzione albero
    - pruning
    - si puo` ridurre l'accuratezza della costruzione dell'albero
    - i dati per il pruning devono riflettere la vera distribuzione dei dati

Estensioni

- Split su attributi multipli
- gestione training set in memoria secondaria
  - tecniche machine learning e statistica in memoria centrale

Clustering

- Dati del problema:
  - base di dati di oggetti
- Problema:
  - trovare una suddivisione degli oggetti in gruppi in modo che:
    - gli oggetti in un gruppo siano molto simili tra di loro
    - oggetti in gruppi diversi siano molto diversi
  - i gruppi possono essere anche sovrapposti o organizzati gerarchicamente

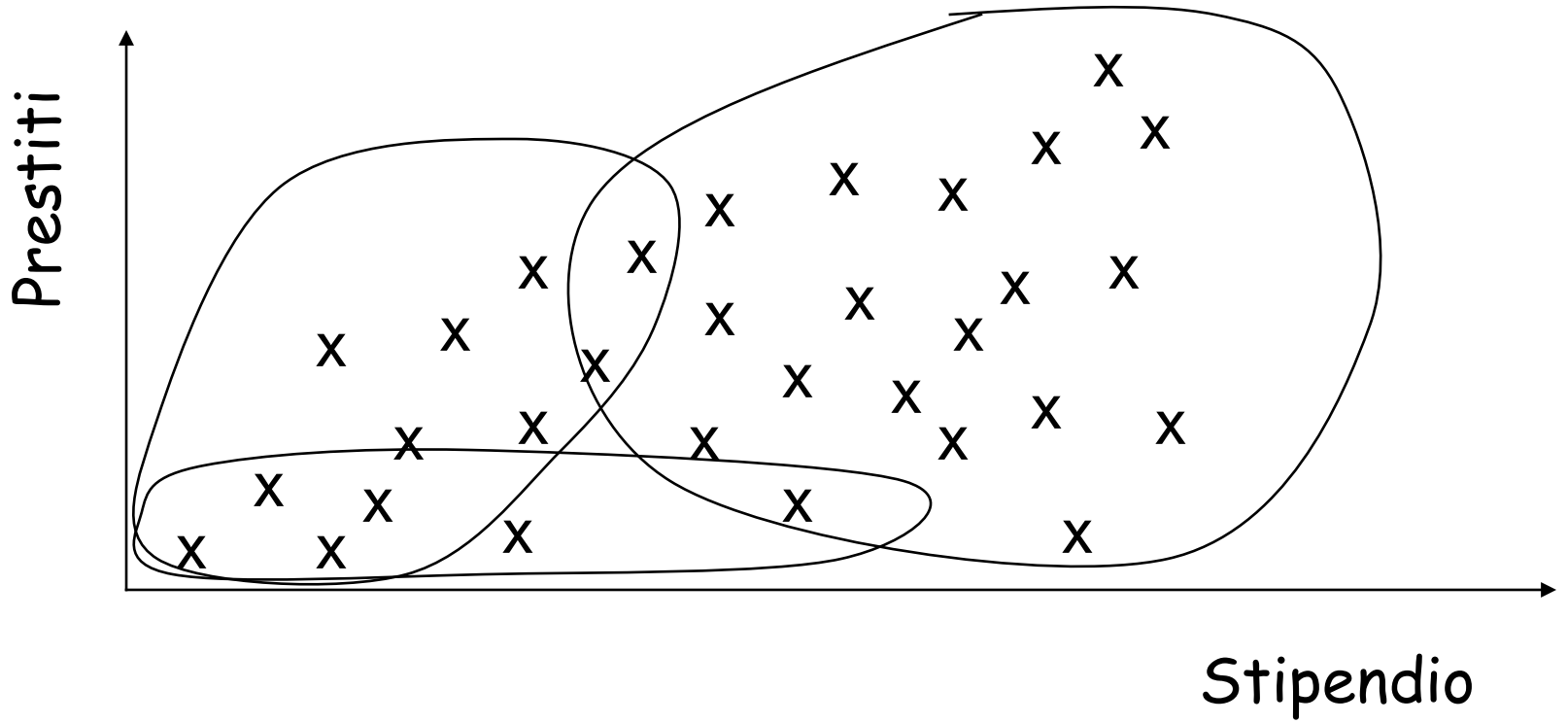
Applicazioni

- Identificazione di popolazioni omogenee di clienti in basi di dati di marketing
- valutazione dei risultati di esperimenti clinici
- monitoraggio dell'attivita` di aziende concorrenti

Settori di sviluppo

- Statistica
- machine learning
- database spaziali
- data mining

Esempio



Approcci

- Approccio statistico:
  - dati = punti nello spazio
  - misure globali
  - oggetti noti a priori
  - non scalabili
- Machine learning:
  - conceptual clustering

Approcci

- Data mining:
  - si determinano i rappresentanti di ogni cluster
  - si cercano gli elementi "simili"
  - si aggiorna il methoid
- Gli algoritmi si differenziano principalmente nella scelta del medoid

Estensioni

- Molti algoritmi solo per memoria centrale
- Utilizzo di tecniche di indice spaziale per clusterizzazione dati su disco

Similarity search

- Dati del problema:
  - base di dati di sequenze temporali
- Problema:determinare
  - sequenze simili ad una sequenza data
  - tutte le coppie di sequenze simili

Applicazioni

- Identificazione delle societa` con comportamento simile di crescita
- determinazione di prodotti con profilo simile di vendita
- identificazione di azioni con andamento simile
- individuazione porzioni onde sismiche non simili per determinare irregolarita` geologiche

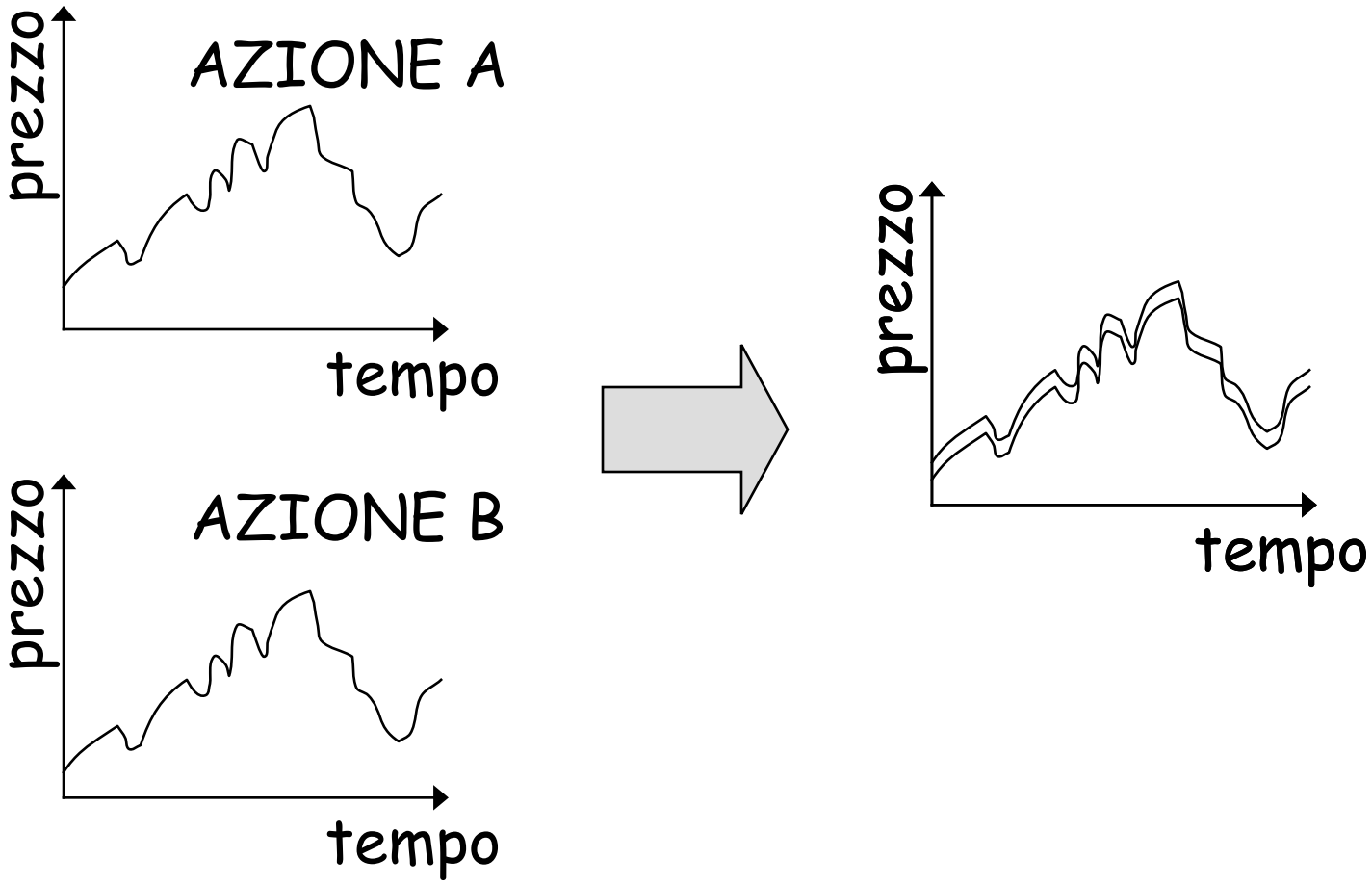
Settori di sviluppo

- Database temporali
- speech recognition techniques
- database spaziali

Tecniche

- Due tipi di interrogazione
  - match completo: la sequenza cercata e le sequenze della base di dati hanno la stessa lunghezza
  - match parziale: la sequenza cercata puo` essere sottosequenza di quelle recuperate dalla base di dati
- Possibilita` di traslazioni, variazioni di scala

Esempio



Misure

- Euclidea
  - match con sequenze aventi la stessa lunghezza
    - si devono generare tutte le sottosequenze aventi la lunghezza della sequenza di query
- correlazione
  - non si devono generare le sottosequenze

Approccio database spaziali

- Ogni sequenza
  - insieme di punti in uno spazio multi-dimensionale

- uso tecniche di indice spaziale per trovare sequenze simili

Data mining

Indicazioni di ricerca

Recenti direzioni di ricerca

- Definizione di linguaggi di alto livello per la specifica di criteri di estrazione della conoscenza
- gestione di tipologie diverse di dati (oggetti, attivi, deduttivi, spaziali, temporali, multimediali, ecc.)
- algoritmi incrementali
- nuove tipologie di applicazioni (web)

Argomenti per seminari

- Applicazione delle tecniche di ILP al data mining
- Data mining spaziale
- Data mining temporale
- WWW data mining