Il processo di KDD

PDF created with pdfFactory trial version www.pdffactory.com

Introduzione

- Crescita notevole degli strumenti e delle tecniche per generare e raccogliere dati (introduzione codici a barre, transazioni economiche tramite carta di credito, dati da satellite o da sensori remoti, servizi on line ...)
- Sviluppo delle tecnologie per l'immagazzinamento dei dati, tecniche di gestione di database e data warehouse, supporti piu' capaci e piu' economici (dischi, CD) hanno consentito l'archiviazione di grosse quantita' di dati

PDF created with pdfFactory trial version www.pdffactory.com

Introduzione

- Simili volumi di dati superano di molto la capacità di analisi dei metodi manuali tradizionali, come le query ad hoc. Tali metodi possono creare report informativi sui dati ma non riescono ad analizzare il contenuto dei report per focalizzarsi sulla conoscenza utile
- Emerge l'esigenza di utilizzare tecniche e strumenti con la capacità di assistere in modo *intelligente* e *automatico* gli utenti decisionali nell'estrazione di elementi di conoscenza dai dati

PDF created with pdfFactory trial version www.pdffactory.com

Introduzione

Queste tecniche e strumenti sono al centro del campo emergente del

Knowledge Discovery in Databases (KDD)

PDF created with pdfFactory trial version www.pdffactory.com

Introduzione

- Il termine ***knowledge discovery in databases*** indica l'intero processo di ricerca di nuova conoscenza dai dati
- Il termine di ***data mining*** si riferisce all'applicazione di algoritmi per estrarre pattern dai dati senza considerare gli ulteriori passi che caratterizzano il processo di KDD (come, ad esempio, incorporare appropriata conoscenza a priori e fornire una opportuna interpretazione dei risultati)

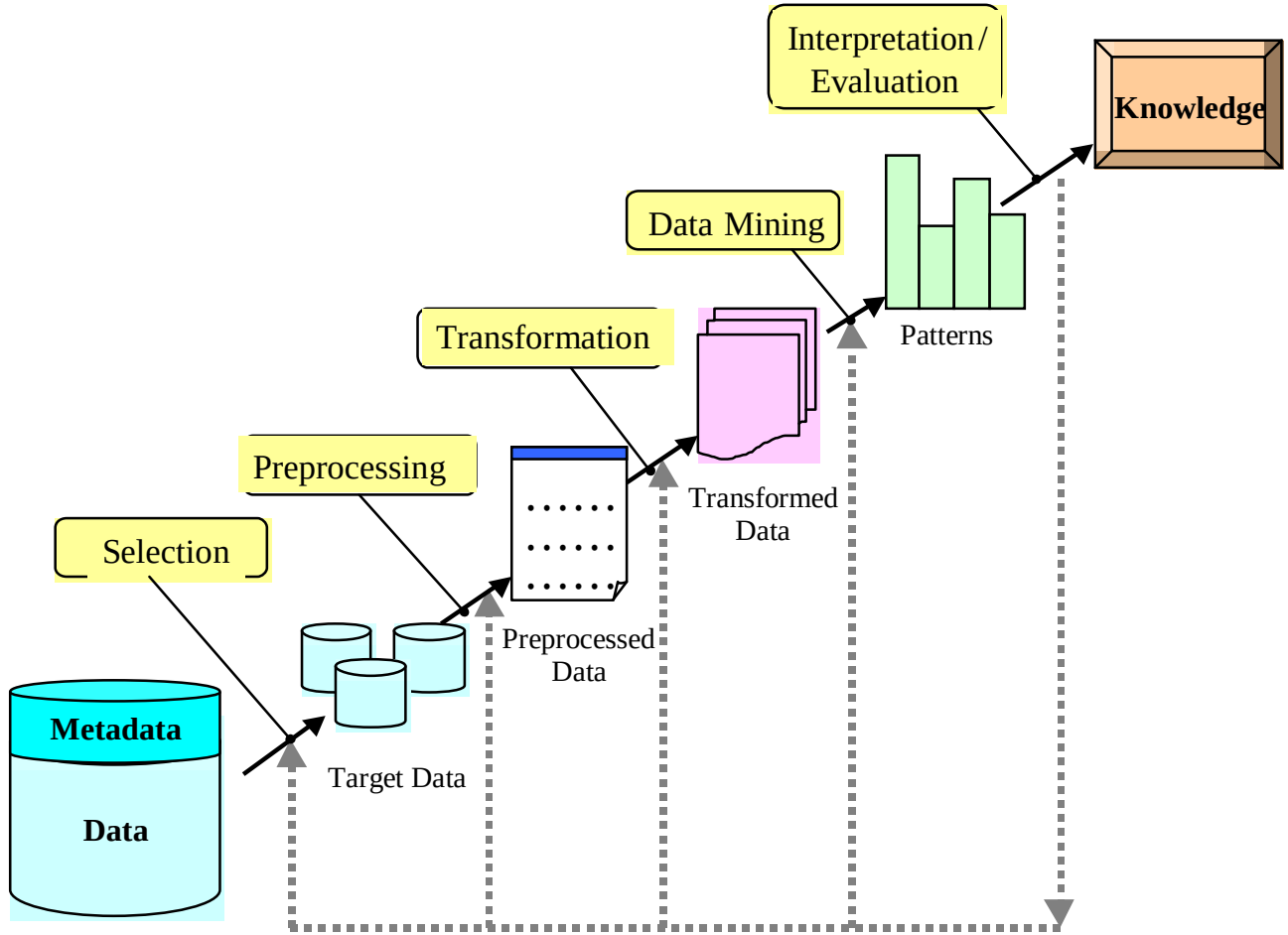
PDF created with pdfFactory trial version www.pdffactory.com

Introduzione

- Pertanto l'intero processo, tipicamente interattivo e iterativo, di ricerca, estrazione ed interpretazione di pattern dai dati, che indichiamo come KDD, coinvolge l'applicazione ripetuta di specifici metodi e algoritmi di data mining e l'interpretazione dei pattern generati da tali algoritmi
- Nel seguito forniremo una definizione più dettagliata di KDD e una panoramica sui metodi e gli algoritmi di data mining più usati

PDF created with pdfFactory trial version www.pdffactory.com

Il processo di KDD (1)



PDF created with pdfFactory trial version www.pdffactory.com

Il processo di KDD: le fasi

1. Sviluppo e approfondimento del dominio di applicazione, della conoscenza disponibile a priori e degli obiettivi dell'utente finale.
2. Creazione di un target data set: selezione del data set o focalizzazione su un sottoinsieme di variabili o di campioni di dati oggetto del processo KDD.

PDF created with pdfFactory trial version www.pdffactory.com

Il processo di KDD: le fasi

3. Cleaning dei dati e preprocessing: operazioni di base come la rimozione del rumore o degli *outliers* se è il caso, raccolta delle informazioni necessarie per modellare o tener conto del rumore, messa a punto di strategie per gestire i dati mancanti e per gestire i dati tempo-varianti.

PDF created with pdfFactory trial version www.pdffactory.com

Il processo di KDD: le fasi

4. Riduzione dei dati e proiezione: rappresentazione dei dati in modo opportuno in relazione agli obiettivi della ricerca. Riduzione delle dimensioni e impiego di metodi di trasformazione per ridurre l'effettivo numero di variabili da sottoporre al processo di ricerca.

PDF created with pdfFactory trial version www.pdffactory.com

Il processo di KDD: le fasi

5. Scelta del compito del processo di data mining: identificazione dell'obiettivo del KDD, stabilire, cioè se si tratti di una classificazione, di una regressione, di un clustering…
6. Scelta dell'algoritmo o degli algoritmi di data mining: selezione dei metodi da usare per ricercare pattern nei dati. Questa fase comprende la decisione su quali modelli e parametri potrebbero essere appropriati e il matching di un particolare metodo di data mining con i criteri generali del processo KDD (per es. l'utente finale potrebbe essere maggiormente interessato alla comprensione del modello piuttosto che alle sue capacità predittive).

PDF created with pdfFactory trial version www.pdffactory.com

Il processo di KDD: le fasi

7. Data mining: ricerca di pattern di interesse in una particolare forma di rappresentazione o su un set di rappresentazioni diverse (regole di classificazione, alberi decisionali, regressione, clustering…). Il risultato del processo di data mining è considerevolmente influenzato dalla correttezza delle fasi precedenti.

PDF created with pdfFactory trial version www.pdffactory.com

Il processo di KDD: le fasi

8. Interpretazione dei pattern trovati e possibile ritorno alle fasi iniziali per ulteriori iterazioni.
9. Consolidamento della conoscenza estratta: incorporazione di tale conoscenza nel sistema di performance o, semplicemente, documentazione e reporting alle parti interessate. Questa fase include anche il controllo per la risoluzione di potenziali contraddizioni con la conoscenza precedentemente disponibile.

PDF created with pdfFactory trial version www.pdffactory.com

Gli algoritmi di Data mining

PDF created with pdfFactory trial version www.pdffactory.com

Differenze tra Data Retrieval e Data Mining

Data Retrieval

- Quanti sono i clienti che hanno età tra 30 e 50 anni e comprano Diet Coke
- Quali documenti contengono la parola “Sanità”
- Quanti brevetti ha depositato la società Colgate nel 1999

Data Mining

- Quali sono le caratteristiche dei miei clienti
- Quali sono gli argomenti trattati da un insieme di documenti
- Quali sono i miei concorrenti e come evolve la loro attività

PDF created with pdfFactory trial version www.pdffactory.com

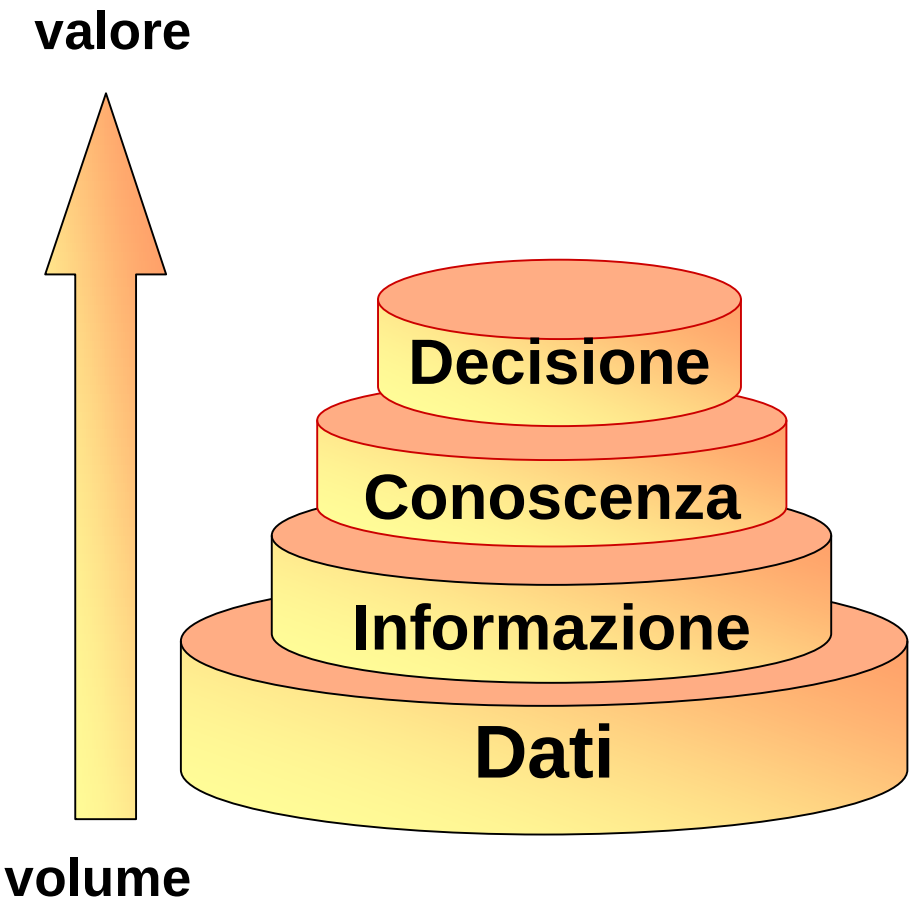
I vantaggi del Data Mining

- Trattamento di dati quantitativi, qualitativi, testuali, immagini e suoni
- Non richiede ipotesi a priori da parte del ricercatore
- Non richiede ipotesi sulla forma distributiva delle variabili
- Possibilità di elaborare un numero elevato di osservazioni
- Possibilità di elaborare un numero elevato di variabili
- Algoritmi ottimizzati per minimizzare il tempo di elaborazione
- Semplicità di interpretazione del risultato
- Visualizzazione dei risultati

PDF created with pdfFactory trial version www.pdffactory.com

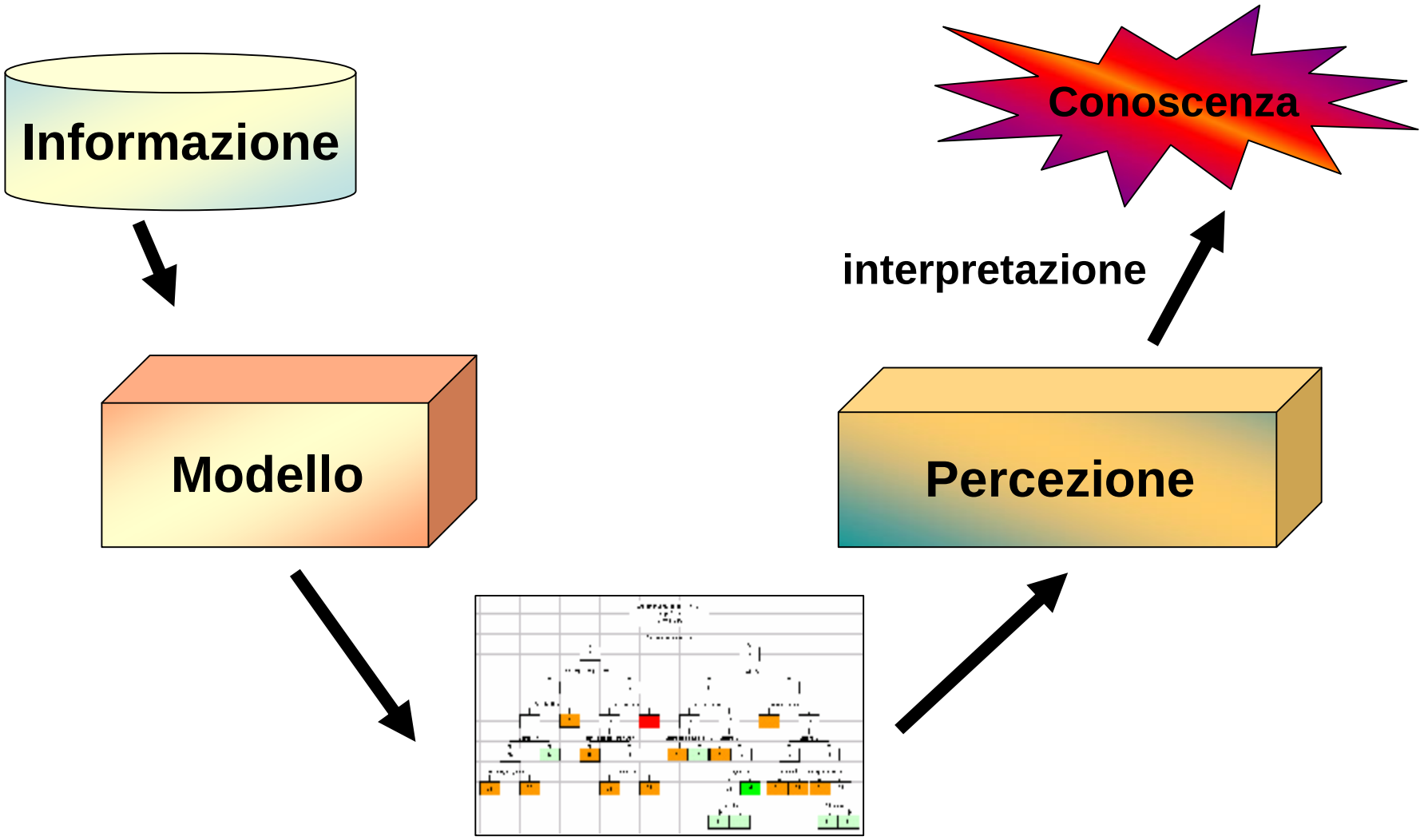
Perché sono necessari strumenti di Data Mining

- Quantità dei dati
- Natura dei dati
- Rapida evoluzione del mercato
- Inadeguatezza degli strumenti tradizionali



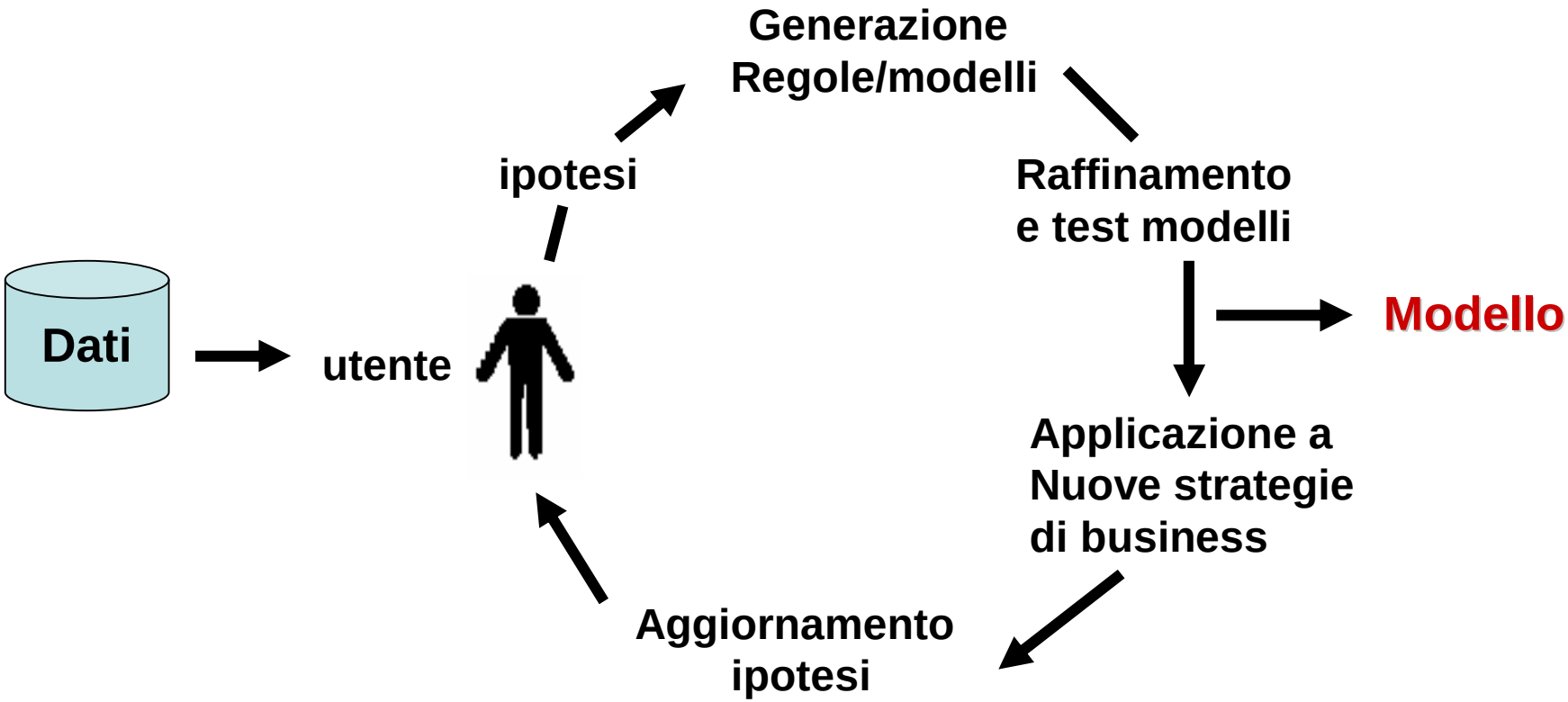
PDF created with pdfFactory trial version www.pdffactory.com

Tecniche di Knowledge Discovery



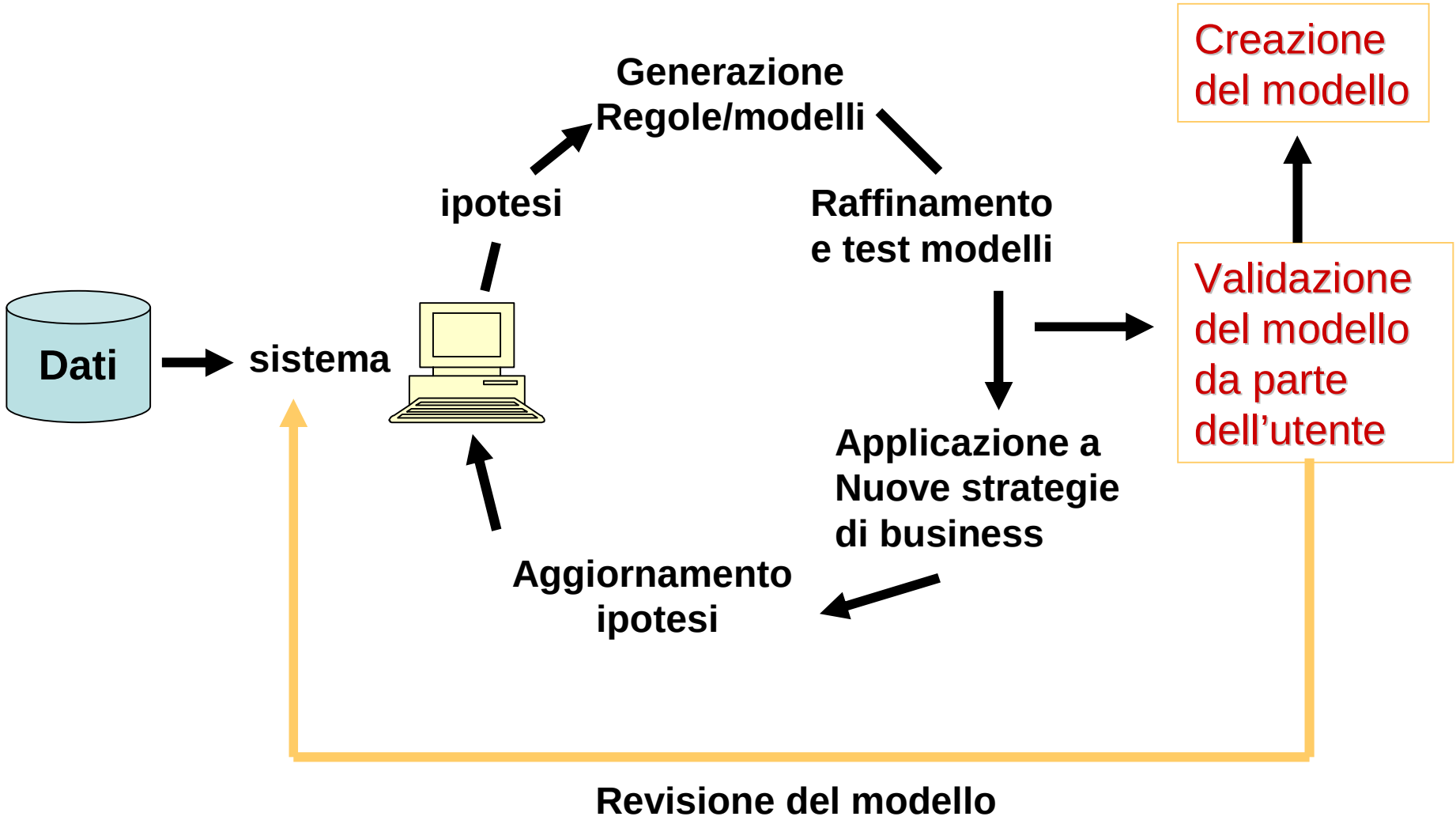
PDF created with pdfFactory trial version www.pdffactory.com

Interazione tra utente e sistema di Data Mining



PDF created with pdfFactory trial version www.pdffactory.com

Generazione del modello a iniziativa del sistema



PDF created with pdfFactory trial version www.pdffactory.com

Problemi di Data Mining

- **L'Analisi per gruppi** suddivide una popolazione in sottoinsiemi disgiunti secondo definiti criteri.
- **La Classificazione** cataloga un fenomeno in una certa classe secondo un insieme di regole predeterminate.
- **Le Regole d'associazione** sono legami di casualità validi tra gli attributi delle osservazioni di un *data set*.

PDF created with pdfFactory trial version www.pdffactory.com

Metodi di Data Mining

- **Esplorazione mediante la visualizzazione multidimensionale** (scaling multimensionale, analisi di regressione logistica, stepwise, analisi delle corrispondenze)
- **Associazione e sequenze** (usate nella market basket analysis per misurare l'affinità dei prodotti)
- **Clustering** (segmentazione della clientela in gruppi omogenei)
- **Analisi Fattoriale** (per determinare il numero di fattori da estrarre)
- **Modelli previsivi**
  - di classificazione (**Alberi di Decisione**)
  - **Reti Neurali**
- **Mappe di Kohonen** (Reti Neurali non supervisionate)
- **Algoritmi Genetici**

PDF created with pdfFactory trial version www.pdffactory.com

Esempi di applicazioni

Esempio	***Tipo di problema***	***Tecnica adottabile***
Quali sono i tre principali motivi che hanno indotto il mio cliente a passare alla concorrenza?	Classificazione	Reti Reurali Decision Tree
Quali sono le fasce di clienti a cui posso offrire nuovi prodotti?	Clustering	Reti Neurali Decision Tree
Quali sono le probabilità che un cliente che ha aperto un c/c acquisterà anche il prodotto x in breve tempo?	Sequencing	Tecniche statistiche Rule induction
Quali sono le probabilità che un cliente acquisti due prodotti completamente differenti?	Associazione	Tecniche statistiche Rule induction
Quale sarà il prezzo del titolo tra un giorno/mese ecc?	Previsione	Reti neurali Tecniche statistiche

PDF created with pdfFactory trial version www.pdffactory.com

Data mining

Ma non se ne occupava la statistica?

J. Kettenring (ex- presidente dell'ASA) definisce la statistica come "*la scienza di apprendere dai dati*"

Tecniche statistiche orientate alla scoperta di strutture di relazione e di modelli

- **Analisi esplorativa**
- **Analisi esplorativa multivariata**
  - Analisi delle componenti principali
  - Analisi delle corrispondenze
  - Analisi dei cluster
  - Ecc.

PDF created with pdfFactory trial version www.pdffactory.com

Data mining

Cosa c'e' di nuovo nel Data mining?

• La possibilità di gestire enormi quantità di dati, che rendono obsoleta la definizione classica di grandi campioni (miliardi di record e terabytes di dati non sono inusuali)

•Le recenti tecniche che provengono dal mondo dell'ingegneria informatica (reti neurali, alberi di decisione, regole di inclusione)

•Interessi commerciali nel valorizzare le informazioni esistenti al fine di proporre soluzioni "individuali" per una determinata categoria di clienti

•Disponibilità di nuovi pacchetti, di facile uso, diretti sia a coloro i quali devono assumere le decisioni che agli analisti (ma molto più costosi!)

PDF created with pdfFactory trial version www.pdffactory.com

Data mining

Il text mining

Potenti tecniche sono disponibili per classificare, analizzare, e raggruppare informazioni o documenti creati con pacchetti di video scrittura

Esempi

•Classificazione delle notizie dei giornali

•Raggruppare e-mail secondo argomenti prestabiliti

•Archiviare in automatico i documenti in base al loro contenuto

•Ecc.

PDF created with pdfFactory trial version www.pdffactory.com

Conclusioni

Il Data mining è una disciplina in grande crescita che si è sviluppata al di fuori della statistica nel mondo dei DBMS, principalmente per motivi commerciali.

Oggi il DM si può considerare come una branca della statistica esplorativa con l'obiettivo di individuare *inattesi* e *utili* modelli e regolarità nei dati mediante l'uso di algoritmi classici e nuovi.

PDF created with pdfFactory trial version www.pdffactory.com

AVVERTENZE ALL'USO

L'espressione *inattesi* non deve essere fuorviante: un ricercatore ha una maggiore possibilità di scoprire qualcosa di interessante se **ha familiarità con i dati.**

L'*utilità* delle regolarità individuate nella struttura dei dati va verificata. **Le associazioni sono solo correlazioni e non implicano relazioni di causa-effetto.**

Non va infine dimenticato che nell'applicazione di questi algoritmi è necessario effettuare valutazioni dell'incertezza e del rischio e pertanto non si può prescindere dall'uso di test per la verifica della validità dei risultati ottenuti (suddividere la base di dati in sotto campioni e verificare se si ottengono gli stessi risultati).

PDF created with pdfFactory trial version www.pdffactory.com