Data Mining: Metodi ed Applicazioni

1. Processo KDD ed elementi di data mining
2. Metodi classificanti
   - regole di decisione
   - alberi di decisione
   - esempio più vicino (nearest neighbor)
   - macchine a supporto vettoriale (SVM)
3. Metodi non supervisionati
   - regole d´associazione
   - analisi di clustering
4. Estrazione di sottogruppi
5. Analisi di web dati
6. Analisi di dati spaziali

Processo KDD ed Elementi di Data Mining

1. Definizioni
2. Tipi di dati
3. Tipi di conoscenze
4. Scoperta di conoscenze
5. Le fasi del processo

KDD: Definizioni

Scoperta di conoscenze in basi dati (KDD)
Processo nontriviale per identificare validi, nuovi, possibilmente utili e comprensibili modelli o regolarità in dati

Data mining
fase d´analisi di dati nel processo KDD

Apprendimento automatico
Modellatura e realizzazione di fenomeni di apprendimento
Compiti di apprendimento: (input, output, restrizioni)
Apprendimento di funzioni da esempi

Esempio d´un Compito di Data Mining

Apprendimento da esempi classificati (classificazione)

Una funzione sconosciuta f associa un valore target agli oggetti d´una popolazione

Oggetti: transazioni di carta di credito
Valore target: fraudolente (sì o no)

Input:
esempi (descrizione d´oggetti: valori per variabili, compreso target)
Output:
descrizione di una funzione approssimata h (ipotesi), stima d´errori
Restrizioni:
ammissibili errori, tempo, memoria

Processo di KDD

Le fasi

Comprendere l´applicazione, stabilire lo scopo
Acquirare ed integrare, pre-elaborare e visualizzare di dati
Scegliere il metodo d´analisi
Creare dati per analisi: campionamento, trasformare, pulire
Selezionare i parametri del metodo, usare metodo
Validare risultati, selezionare e visualizzare risultati importanti
Utilizzare risultati: riporti, sistemi operativi
Convalidare la soluzione in pratica

Il processo reitera con molte decisioni dell´analista umano

Scoperta di Conoscenze in Basi Dati

Nessun necessità per una definizione parafrasata:

Conoscenza verità espressa e giustificata su un dominio rappresentata con un linguaggio formale

Scoperta autonoma generazione e verifica delle ipotesi

Basi dati ben strutturati e mantenuti depositi di dati su dominio di mondo reale

Dimensioni per Classificare Dati

Main data type:	observation / transaction	textual	multimedia	
Representativity:	complete population	sample of convenience	random sample	stratified sample
Variable type:	binary	categorical	continuous	mixed
Missing data:	no	yes		
Conclusiveness:	low	medium	high	
Size:	moderate	large	very large	vast
Dimensionality:	low	medium	high	
Dynamics:	static	timely evolving		
Distribution:	local	fixed locations	scattered on net	
Object heterogeneity:	one object class	multi-valued attributes	multiple object classes	
Time reference:	one cross section	2 independent cross sections	series of independent cross sections	longitudinal data
	continuous	time series		
Space reference:	point	line	area	surface
Text structure:	unstructured	structured parts	hypertext	
Text languages:	English	other languages	mixed collection	multilingual text
Text quality:	high	low (e.g. email)		
Text size:	full size	abstract		
Hybrid forms:	no	mixed observational data	observational & text & multimedia	
Aggregation:	micro data	macro data		
Meta data:	no	data dictionary	domain knowledge	

Tipi di Basi Dati

Relazionali
Orientato agli oggetti
Multidimensionali / OLAP
Deduttivi
Paralleli
Distribuiti

Formalismi di Conoscenze

Tabella di contingenza
Pattern di sottogruppo
Regola
Albero di decisione
Cluster
Rete probabilistica
Relazione funzionale
Rete neuronale
Tassonomia

Tabella di Contingenza

Tabella con le **frequenze** congiunte di due **variabili**

età contro sesso:

età/ sesso	1-6	7-15	16-20	21-45	45-60	60-75	75+
maschio	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}
femminile	A_{21}	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}

A_{ij} valore realizzato con $A=a_i$ e $B=b_j$

E_{ij} valore atteso con $A=a_i$ e $B=b_j$ $E_{ij} = f(A=a_i) \times f(B=b_j) / N$

Tabella di Contingenza

Per manifestare relazione statistiche fra (2) variabili (categoriche)
Tipo di conoscenza molto elementare

Tabella di contingenza **interessante**:
In che grado A_{ij} è diverso da E_{ij} per non più poter pretendere
che E_{ij} sia la distribuzione vera
(ipotesi nulla)

Chi quadrato = $\sum_{i,j} (A_{ij} - E_{ij})^2 / E_{ij}$ i=1, …, I j=1, …, J

Cramer´s V = $\sqrt{}$ (chi2 / (N x min(I-1,J-1)))

Significatività statistica (o P-value associato)
misura la consistenza dei dati osservati con l'**ipotesi nulla** formulata,
e quindi la forza dell'evidenza contro la stessa ipotesi

Sottogruppo

Sottogruppo = sottoinsieme di una popolazione
definito con un linguaggio d´interrogazione

Selettore: clausola condizionale $A \in V$ $V \subset$ Dominio(A)
forma congiuntiva di selettori
Sesso=maschio & età=18-25

Estensioni: **interni disgiuntivi**: Nazionalità = austriaca | svizzera
tassonomie: Nazionalità = europea

Interrogazioni multirelazionali:
persona. età=1-6, vicino(persona,impianto-nucleare),
impianto. età =vecchio

Pattern di Sottogruppo

Pattern di sottogruppo
descrive un sottogruppo con una statistica **significativa**

Statistica significativa (di una variabile target nel sottogruppo):
Tasso di un valore di una variabile discreta
media di una variabile continua

Esempi:

Tasso di disoccupazione più alto per giovani maschi con bassa educazione

Giovani povere donne sono infettate con AIDS più dei corrispondenti maschi

La mortalità di cancro di polmone è aumentata per donne durante gli ultimi 5 anni

Regola

If S then T (oppure $S \rightarrow T$) S, T sottogruppi

Criteri elementari di valutazione:

Precisione (confidenza): $|S \cap T| / |S|$
= probabilità condizionale $P(T|S)$

Supporto: $|S \cap T| / N$ (oppure $|S| / N$)

N
$p_0 = |T| / N$

	T	Non T
S	VP Veri positivi	FP Falsi positivi
Non S	FN Falsi negativi	VN Veri negativi

Tipi di Regole

$S \rightarrow T$

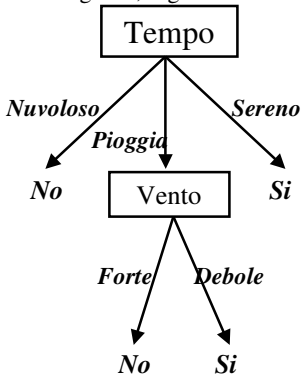
Regole di classificazione
T è fisso, scoperta di un sistema di regole S_i per predire T
spesso è data una variabile target (t), T : $t=t_0$, T_i: $t=t_{i\ i=1,\ldots,k}$

Regole caratteristiche
S è fisso, scoperta di un sistema di regole T_i per descrivere le proprietà di S

Regole di associazione
attributi binari A_i che descrivono una transazione
ogni attributo rappresenta un articolo, evento; 1 significa presenza nella stessa transazione:
articolo comprato da un cliente, parola presente in testo
S, T definiti dai selettori $A_i=1$ (forma congiuntiva), $S \cap T = \emptyset$; S,T insiemi di prodotti, …
insiemi frequente: insieme di transazioni con supporto minimo

Alberi di Decisione

Struttura ad albero di sottogruppi
Successori di sottogruppo:
tutti i selettori con valori d´un solo attributo
valori discreti (intervalli per attributi ordinali)
Successori disgiunti, foglie formano una spartizione



Esempio	Tempo	Temper.	Umidità	Vento	SPORT
1	Sereno	Caldo	Normale	Forte	Si
2	Sereno	Caldo	Alta	Forte	Si
3	Nuvoloso	Freddo	Alta	Forte	No
4	Sereno	Caldo	Alta	Debole	Si
5	Pioggia	Freddo	Alta	Debole	Si
6	Pioggia	Freddo	Alta	Forte	No

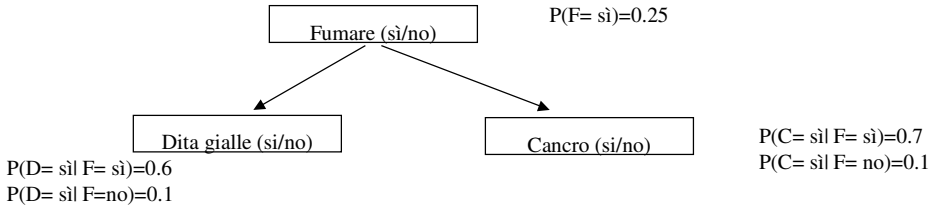
Clustering

Non esistono attributi target
Il metodo deve scoprire le classi, raggruppando oggetti simili
(=con valori simili degli attributi) nella stessa classe
Misura di similarità (distanza) fra oggetti
vettori di *d* dimensioni: differenze dei vettori
utilità dipende dall´applicazione
Forma: elongata / sferica simile ad un membro / tutti i membri
- Spartizione di classi disgiunte oppure
- Membro fa parte di diversi classi (con probabilità individuale)
- Sistema gerarchico (ad albero) di cluster

Usato come
- descrizione sommaria
- fondamento per modelli predittivi più accurati

Rete Probabilistica (Bayesiana)

Grafo diretto aciclico
Ogni nodo è associato ad un´unica variabile
Ogni arco x→y rappresenta la dipendenza condizionale fra i due nodi
variazione di x produce variazione di y con valori fissati di nodi non discendenti di y (struttura causale)
Insiemi di distribuzioni locali probabilità condizionale di parenti: P(C|F),P(D|F)
Per predire la probabilità di ogni variabile condizionale di valori di un insieme di altri variabile: P(C=sì |D=no)=0.187, P(C=sì)=0.25



Indipendenza Condizionale

X,Y,Z variabili discrete

X è indipendente di Y, dato Z :: $P(X=x_i|Y=y_j,Z=z_k)=P(X=x_i|Z=z_k)$

$i=1,\ldots I \quad j=1,\ldots,J \quad k=1,\ldots,K \quad (Dominio(X)=x_1,\ldots,x_I, \ldots)$

La distribuzione probabilistica di X è indipendente del valore di Y dato un valore di Z

Ogni nodo è indipendente dei suoi nondiscendenti, dato (condizionale di) i suoi parenti immediati

Cancro è indipendente di Dita gialle, dato Fumare

Rete Probabilistica: Compiti

- **Inferenza**
  predire la probabilità di una variabile
  condizionale dei valori di un insieme di altre variabili

- **Apprendimento dei parametri (probabilità condizionali),**
  dato la struttura

- **Apprendimento della struttura**

Tassonomia

Gerarchia di concetti

Spartizione gerarchica degli oggetti

Medicina, Biologia, etc.

Clustering, (albero di decisione)

Spartizione dei valori di un solo attributo

Ad ogni nodo sono associate definizione degli oggetti che formano la classe rappresentata dal nodo, e proprietà derivate della classe

Relazioni Funzionali

Relazioni funzionali matematiche
Relazioni funzionali empiriche (errori, oggetti mancanti)
Equazioni (scoperta scientifica automatica)
Regressioni (statistica)
Equazioni in sottogruppi
Alberi di regressione
Ingegneria, Scienza

Reti neuronali

Pattern: Nugget versus Modello

Nugget: pattern descrivendo una parte della populazione
- sottogruppo
- regola

Modello: pattern descrivendo tutta la populazione
- tabella di contingenza
- albero di decisione
- clustering
- rete probabilistica
- relazione funzionale
- tassonomia

Ma: un sistema di sottogruppi, regole
una cellula di una tabella, foglia di albero, un cluster

Scoperta di Conoscenze

Compiti generali

Spazi di ricerca

Validazioni

Restrizioni

Compiti / Scopi di Usare Conoscenze

Classificare

predire una proprietà di un caso futuro / sconosciuto
valore discreto: cliente è fraudolente
valore continuo: spese di un cliente (regressione)

Descrivere / Riassumere un dominio

presentare evoluzioni di variabili e dependenze fra variabili
sviluppo di prezzi di mercato, vendite, quotazioni di borsa

Esplorare dependenze fra variabili

analizzare correlazioni, causalità

Ottimizzare

soluzione migliore d´un problema combinatorio
trovare i parametri ottimali d´un processo di produzione

Spazi di Ricerca

Spazi delle ipotesi (patterns, modelli, regole)
Ipotesi: speculazione, congettura di conoscenza possibile
che possa essere supportata dai dati
Necessario verificare e perfezionare le ipotesi:
iterativo processo di generare e controllare ipotesi
Rappresentazione, valutazione, e **ricerca**
Ricerca: trovare un insieme d´ipotesi adatte
Problema combinatorio in uno spazio parzialmente ordinato
ipotesi più specifica

Ipotesi Parzialmente Ordinate

Subgruppo: Maschio → Maschio e Giovane

Regola: if Maschio then Incidente → if Maschio e Giovane then Incidente

Albero: Albero → Albero con uno split aggiunto ad una foglia

Equazione: Equazione → Equazione con una variabile (termine) aggiunta

Metodi di Ricerca

Ricerca esauriente, completa:
enumerare ipotesi, potare ipotesi inferiori alle più migliori già trovate

Ricerca locale (in vicinanze): operatori generando successori / vicini
trovare vicini che sono migliori

Ricerca gradiente: determinare i parametri dell´ipotesi minimizzando gli errori (gradiente dell´errore)

Ricerca stocastica: selezione stocastica d´operatori

Errori

Classificazione / Apprendimento da esempi

P populazione d´oggetti
$x=x_1,...,x_k$ variabili indipendenti y variabile target
D Distribuzione probabilistica D(x,y) su P
H insieme di funzioni ammissibili (linguaggio d´ipotesi)

Dato:
E insieme d´esempi (x,y) con y=f(x) per una funzione sconosciuta f
Trova:
$h \in H$ $errore_D(h,f)$ minimale

D ed *errore vero* $errore_D(h,f)$ sono sconusciuti

Errore d´Addestramento (training error)

Training error

$errore_E(h) := \sum_{(x,y)\in E} errore(h(x),y)$

y variabile discreta: errore(h(x),y):=0, se h(x)=y, altrimenti 1

Y variabile continua: $errore(h(x),y):=(h(x)-y)^2$

y binaria	y=1	y=0
h(x)=1	VP Veri positivi	FP Falsi positivi
h(x)=0	FN Falsi negativi	VN Veri negativi

Statistiche:
Percentuale di misclassificati esempi: (FP+FN)/N
Tasso falso positivo: FP/(FP+VN)
Chi2 …
Funzione di perdita (loss)
0-1 perdita, tipi di misclassificazione (FP vs. FN)

Insieme di Validazione

Test error (T un insieme di validazione $T \cap E = \emptyset$)

$errore_T(h) := \sum_{(x,y)\in T} errore(h(x),y)$

L´errore d´addestramento: come approssima h gli oggetti di E

L´errore di validazione è una stima per l´errore vero

E esempi con valori target
A insieme d´addestramento $A \subset E$, $|A| = 2/3\ |E|$
T insieme di validazione = $E \setminus A$

Un solo insieme di validazione basta, se ci sono molti esempi

Cross Validazione

k-fold:

$E = E_1 \cup ... \cup E_k \quad E_i \cap E_j = \emptyset \quad |E_i| = |E_j|$

h_i funzione appresa su $E \setminus E_i$

$Errore_{CV} = \sum errore_{E_i}(h_i) / k$

Stima per errore vero

k=3,...,10 dipende da |E|

Cross Validazione e Algoritmi Greedy

Per stimare statistiche, ottimizzare parametri

Incluso nella ricerca, ad esempio:

Algoritmi greedy scovano uno spazio di componenti da aggiungere ad un modello corrente (split ad un albero, selettore congiuntivo, variabile)

Procedura di multipli paragoni:

Costruire multipli items modelli, componenti, parametri $m_1,...,m_n$
Stimare score $s_i=f(m_i,E)$ E insieme d´addestramento
Selezionare item con massimo score

Cross validazione può migliorare gli scores

Mà: tempo x k

Problemi di Multipli Paragoni

Overfitting

Oversearching

Errori di feature selection

Overfitting

Una componente aggiuntata aumenta la precisione di un modello sui dati d´addestramento, ma la precisione vera non viene aumentata

Distribuzione statistica per il massimo è differente

Oversearching

Una ricerca in uno spazio più ampio aumenta la precisione di un modello sui dati d´addestramento, ma la precisione vera non viene aumentata

Distribuzione statistica per il massimo dipende dal numero di componenti

Errori di Feature Selection

Variabili con molti valori vengono scelti per uno split

Distribuzione statistica per il massimo dipende dal numero di componenti

Restrizioni sui Spazi di Ricerca

Per limitare la ricerca ed escludere risultati irragionevoli

Applicati durante o dopo la ricerca

Tipi di restrizioni

- **Risorsi: tempo, memoria utilizzabile**
- **Sintattico: linguaggio d´ipotesi**
- **Dominio: ad esempio gruppi di variabili simili**
- **Qualità: precisione, supporto, e cosi via**
- **Ridundanza: ipotesi più specifica non è migliore**

Le Fasi del Processo

Comprendere l´applicazione

conoscenze rilevante esistente
scopo dell´utente
terminologia
esigenze di qualità, criteri di successo, costi-benefitti
risorsi
programma di compimento
sicurezza, privacy
aspetti legali

Applicazioni

Analisi di mercato
comportamento di clienti, sales/direct marketing campains

Scoperta di frode **transazioni bancari**

Analisi di rischio **meriti di crediti, investimenti**

Controllo della produzione **ottimi parametri della produzione**

Fault analysis **difetti in reti**

Text/web/multimedia mining

Scienze **medicina, farmacologia, biologia, ambiente**

Integrare Dati per Analisi

Raccogliere e load di dati
Che tabelle sono disponibili?
Che attributi sono disponibili, rilevanti?
Come congiungere le tabelle?
Definire il modello di dati
Generare (separata) base di dati
Verificare qualità di dati
Rappresentativi dati?
Che dati mancano?
Popolazione target simile ai dati disponibili?
Precisione di dati?
Esplorare i dati Visualizzazioni e statistiche
Statistiche fondamentale: distribuzioni e correlazioni di variabili
Che trasformazioni?

Visualizzare Dati

Grafiche statistiche interattive
bar chart, histogram, scatter plot
interazioni con grafiche (interrogazioni, zooming)
paradigma di linked views
Tecniche interattive multivariate per molti variabili (>20)
parallel coordinates
Animazioni (tempo, spazio, p.e. traffico in reti)
GIS (Geographic Information Systems)

Visualizzare il processo di un data mining compito
Clementine/SPSS

Selezionare Obiettivo
d´un compito di data mining

Classificazione? Regressione? Clustering? Analisi del trend?

Un problema del mondo reale deve essere armonizzato con un compito astratto

Arte: Semplificare problema,
ma rimuovere solo irrilevanti caratteristiche
Molta esperienza necessario.
P.e. scoperta di frode: classificazione o clustering se non molti esempi positivi sono disponibili?
Importante rappresentare tempo e spazio?

Selezionare Metodo
d´un compito di data mining

Criteri d´applicazione
Criteri di dati

attributi simbolici, categoriche -> alberi, regole
attributi numerici -> NN, SVM, regressioni
nuggets -> sottogruppi
comprensibilità -> regole, sottogruppi

restrizioni di tempo, spazio, ...

Selezionare automaticamente i metodi:
approcci di sistemi esperti versus basato sul caso (NN)

Nessun algoritmo è il migliore per ogni caso

Creare Dati per Analisi

Campionamento troppi casi (> 1 milione)
Campione casuale variabilità di stima spesso inversamente proporzionale al numero
Campione sistematico, strati: gruppi piccoli sopra-rappresentati, pesi
Feature Selection troppi variabili (>100)
Anche per aumentare precisione, semplificare risultati, cancellare variabili irrelevanti
Feature Generation trasformazioni di variabili
Congiungere tabelle e aggregare
Discretizzazione di variabili continui
Pulire errori ed inconsistenze
valori mancanti
omettere casi, derivare valori, tenere come valore speciale

Selezionare Parametri, Usare Metodo

Strategie d´analisi
richiede esperienza
sistemi esperti?

Esempi nelle lezioni seguenti

Validare Risultati, Selezionare e Visualizzare Risultati Importanti

Applicazione di metodi di data mining produce solo elementi di conoscenze

Elementi devono essere verificati e ben organizzati

Verificare contro dati: p.e. cross validazione

Scrutinare contro dominio

- **Risultati inesplicabili?**
- **Difetti di dati: Manca la rappresentatività?**
- **p.e. inchieste: interviste truccate**

Bias del metodo?

- **Applicare diversi alternativi metodi della stessa categoria**
- **Generare diversi tipi di conoscenze (regole, alberi, reti)**
- **Decidere su risultati adatti**

Superare Diluvio di Troppe Ipotesi Verificati

Brute force data mining produce troppe regole, equazioni, ...

- filtrare
- sopprimere
- ordinare
- potare ipotesi piu semplice si adatta meglio su test dati
- combinare e generalizzare
- agglomerare e clustering

Visualizzare Risultati

- **ridisporre fuoco di ricerca**

- **presentare informazioni su sensitività, confidenza, supporto**

- **immediatamente capire perché l´ipotesi è interessante**

- **evitare distrazioni e irrilevanti proprietà**

Utilizzare Risultati: Riporti/Sistemi Operativi

P.e. applicare regole di classificazione if A=a then C=c

Spesso possibilità ristretta di manipolare attributi
non poter cambiare età, ma aspettare che cliente diventa piu anziano e rientra in una categoria piu profittevole

Ragioni legali escludono attributi per selezionare
razzo, sesso, età predicono rendimenti inferiori
offrire assistenza speciale per questi gruppi

Cambiare valore di A (A=a di un cliente)
relazione causale? Verificare quando dati su questi manipolazioni sono disponibili

Ma: Incrementare vendite d´un prodotto può diminuire questi d´un altro prodotto più profittevole

Più ampia prospettiva necessario, spesso solo applicata creativamente dai data owners

Convalidare la Soluzione in Pratica

Verificare le azioni adottate in base a risoltati di data mining

Verificare contro obiettivi d´impresa, criteri di successo

p.e.
Implementazione d´una nuova strategia di marketing:

----> payback time < 1 anno
----> aumento del profitto > 10%