Data Mining e Bioimmagini

Tecniche e applicazioni del Data Mining, classificazione e clustering

Prof. Pierangelo Veltri – 03/010/2023- Autori: Panarello, Saeedzadeh - Revisionatori: Panarello, Saeedzadeh

ATTIVITÀ TIPICHE DEL DATA MINING

- **Classificazione [Predittiva]** si hanno le classi e voglio spiegarle e descriverle in termine di caratteristiche
- **Clustering [Descrittiva]** - scoprire gruppi di dati simili
- **Ricerca di regole associative [Descrittiva]**
- **Ricerca di pattern sequenziali [Descrittiva]**
- **Regressione [Predittiva]** descrivere le forme (funzione) che approssima i dati
- **Individuazione di deviazioni [Predittiva]**

Meccanismo di profilazione: costruisce il profilo (tramite il data mining) di un dato soggetto.

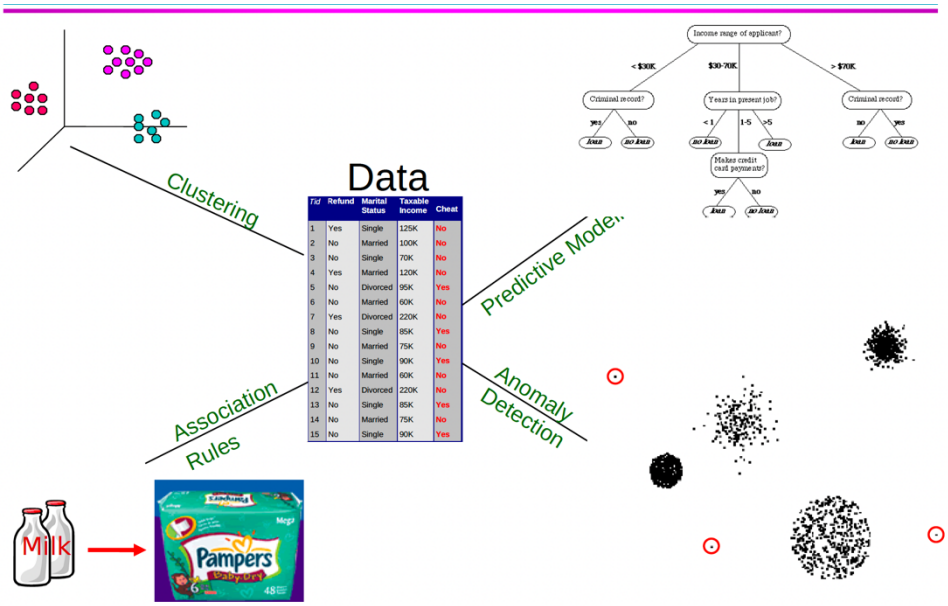
I **task(compiti)** sono generalmente divisi in 2 categorie:

- **Predittivi**: l'obiettivo è di predire il valore di un particolare attributo basato sul valore di altri attributi. Tale attributo d'interesse è detto **target** o **variabile dipendente**, mentre gli attributi usati per creare la predizione sono conosciuti come **esplicativi** o **variabili indipendenti**.
- **Descrittivi:** l'obiettivo è di dedurre patterns (correlazioni, tendenze, clusters, anomalie) che riassumono le relazioni tra i dati. Il data mining descrittivo prevede una fase di post-processing per validare e spiegare i risultati (si ottengono dati più comprensibili e facilmente utilizzabili per analisi successive).

Esempio di costruzione della tabella di rappresentazione dei dati

Si hanno 10 cartelle cliniche di una serie di pazienti oncologici; si ricerca all'interno del dataset uno specifico termine/parola chiave come "carcinoma alla mammella" (tale elemento è la caratteristica di interesse) e si costruisce una tabella: avremo 10 righe (una per ogni cartella) e per colonne si scelgono uno o più descrittori rispetto alla ricerca effettuata precedentemente (come lo stadio del tumore, la presenza o meno di recidive, terapia, ecc.). Nell'incrocio tra riga e colonna si inserisce la frequenza con cui tale termine (il descrittore) risulta presente nella cartella clinica indicata; il risultato che si ottiene è una tabellina di rappresentazione dei dati che può essere usata per fare delle analisi predittive, di profilazione o di clustering.

Nell'immagine si osserva un altro esempio di **tabella di rappresentazione** di dati riguardanti alcuni aspetti economici di un campione. In particolare, nelle colonne sono indicati come descrittori l'effettuazione o meno di rimborsi, lo stato civile, il reddito imponibile e l'aver commesso o meno imbrogli. Tale set di dati, come indicato in figura, può essere utilizzato per successive analisi di natura diverse:



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

- Si può effettuare un **clustering** (raggruppamento) per alcune specifiche caratteristiche, ad esempio per il reddito;
- Si può creare un **modello predittivo** utilizzabile, ad esempio, da una banca per scegliere se concedere o meno un prestito al soggetto in esame;
- Si possono individuare delle anomalie (**anomaly detection**) che possono essere utilizzati per individuare eventuali frodi (si vede più avanti);
- Si possono dedurre delle correlazioni tra caratteristiche tramite **Association Rule (**come sapere che chi compra latte generalmente compra anche pannolini, cosicché quando si organizza la disposizione dei reparti nel supermercato tali prodotti sono in scaffali adiacenti).

In ambito medico, delle possibili applicazioni possono essere:
- L'organizzazione degli strumenti e dei farmaci nei reparti in base a regole associative (ad esempio tenere nello stesso "scaffale" i farmaci necessari per l'intervento farmacologico contro l'infarto del miocardio);
- L'anomaly detection nel follow up di pazienti oncologici per stabilire la periodicità delle visite (si pensi al caso in cui un paziente con cancro ai polmoni sia fumatore, in quel caso un follow up di 6 mesi viene indicato come anomalo in quanto andrebbe ravvicinato per ridurre il rischio di recidive);
- Si può fare clustering dei pazienti in base a specifici sintomi, caratteristiche, tipo di terapia, ecc.

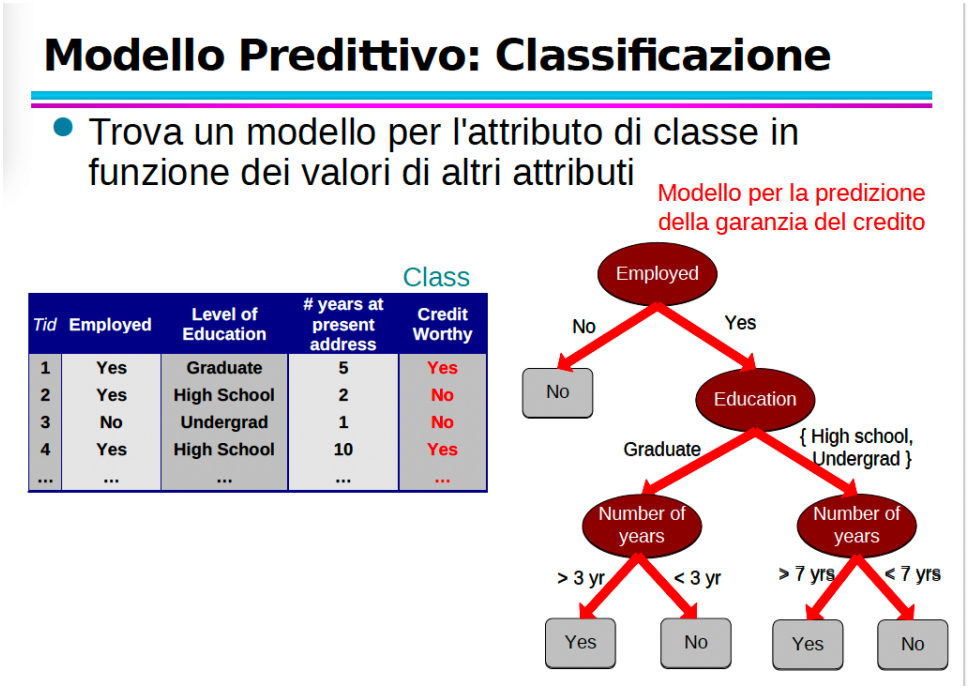
Per fare il clustering è necessario identificare gli attributi d'interesse.

CLASSIFICAZIONE
Data una collezione di record (**training set**), composto da un insieme di **attributi** di cui uno esprime la classe di appartenenza del record, si trova un **modello** per l'attributo di classe che esprima il valore dell'attributo in funzione dei valori degli altri attributi. L'obiettivo è quello di assegnare i **record non noti ad una classe nel modo più accurato possibile.** Viene utilizzato un test set per determinare l'accuratezza del modello. Normalmente, il data set fornito è suddiviso in training set e test set. Il primo è utilizzato per costruire il modello, il secondo per validarlo.

#Esempi
- **Classificazione di un paziente oncologico:** in base alle caratteristiche del tumore (stadio, classificazione TNM, tipo, possibilità di terapie, patologie pregresse, recidiva) si può assegnare al paziente una stima di sopravvivenza.
- **Pazienti sani e malati**: dato un insieme di pazienti descritti da features (ghiandola prostatica, età, proteine...) e da una classe ("sano", "malato") identificare: (l) appartenenza a una delle due classi guardando soltanto una combinazione di features.
- **Analisi dei segnali vocali**: dato un insieme di segnali vocali, acquisiti da pazienti e da soggetti sani, decidere se un nuovo individuo appartiene alla classe sano/malato usando solo il segnale vocale. Dato un insieme di segnali vocali associati a pazienti malati di SLA e pazienti neurologici, identificare la patologia o la tendenza a svilupparla.
- **Predire la struttura di una proteina**: si utilizza un modello che tramite il "**guessing**" (indovina sulla base delle sue conoscenze) stabilisce la possibile sequenza amminoacidica; altra tecnica applicabile è la **meta-prediction** in cui si prendono una serie di modelli a cui si richiede lo stesso task per poi confrontare, tramite una metrica, i risultati ottenuti fino ad avere la sequenza amminoacidi ricercata. Spesso si applica anche sulla predizione del folding delle proteine.

MODELLO PREDITTIVO

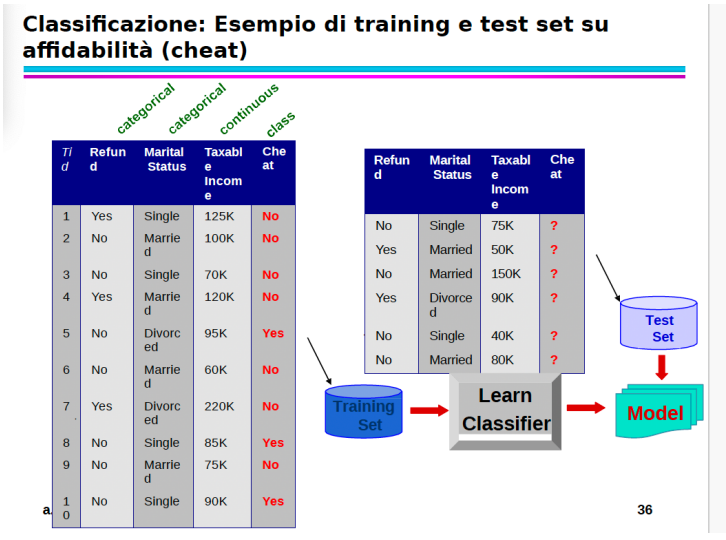
Un modello predittivo è essenzialmente una tecnica (algoritmo) in grado di individuare schemi e andamenti ricorrenti nei dati a disposizione e fornire una previsione di alcune grandezze di interesse. Tramite un modello di classificazione ad albero decisionale è possibile effettuare una classificazione degli individui e raggrupparli in base a caratteristiche d'interesse.



Per esempio, è possibile scegliere in base al tipo e dimensione di ostruzione coronarica l'intervento più adatto per trattare un paziente con IM (trattamento domestico, trattamento in emodinamica, trattamento in cardiochirurgia).

Nell'immagine a sinistra si ha un esempio in ambito finanziario in cui tramite un albero decisionale si "misura" l'affidabilità dei vari candidati per un prestito.

La scelta degli individui della popolazione oggetto d'analisi è importante per aumentare l'affidabilità e l'attendibilità dei risultati. Per esempio, se l'intero dataset è composto da pazienti si hanno risultati diversi da quelli ottenuti prendendo casualmente "per strada" gli individui del dataset (partendo dal presupposto che chi va dal medico ha qualche sintomo/patologia che lo spinge a recarsi dal professionista). Allo stesso modo, se si selezionano per l'analisi del TSH individui che vivono tutti sul mare, si ha una modificazione del risultato rispetto al caso realistico di individui che vivono in aree geografiche diversificate (mare, montagna, città, ecc.). Inoltre, è importante che il dataset di test set non sia lo stesso usato per istruire il modello (training set) altrimenti si avrebbe un risultato di affidabilità del modello falsata.



Costruire un modello significa istruirlo sui vari casi e testarlo sulla base di dati noti.

Una volta applicata la classificazione, è importante verificare la validità ed affidabilità della classificazione riportata. Per far ciò è necessario attribuire una **metrica** che permetta di valutare di quanto il valore ottenuto dal modello di predizione si distanzia dalla realtà: accuratezza, z-score, …

Test basato sull'auto encoder: dato un ECG non noto tira fuori un'informazione relativa alla presenza o meno di anomalie. Il dottorando ha creato dei dataset ed ha affermato che tale sistema ha un errore del 98%. Successivamente ha confrontato il suo sistema con un altro algoritmo

inglese cui affidabilità era del 97,8% e, usando la metrica, ha potuto stabilire la differenza di affidabilità dello 0,2%.

Esempio sull'individuazione di frodi
- Obiettivo: predire l'utilizzo fraudolento delle carte di credito
- Approccio:
• Utilizza le precedenti transazioni e le informazioni sui loro possessori come attributi
– Quando compra l'utente, cosa compra, paga con ritardo, ecc.
• Etichetta le precedenti transazioni come fraudolenti o lecite
• Questa informazione rappresenta l'attributo di classificazione
• Costruisci un modello per le due classi di transazioni
• Utilizza il modello per individuare comportamenti fraudolenti delle prossime transazioni relative a una specifica carta di credito (si ricerca la frequenza degli **item anomali**)

Ad esempio, se su una carta in cui la media delle transazioni è di 300 euro si ha una transazione a di 30000 euro, questa viene segnalata dall'algoritmo come anomala e riferita al titolare della banca, che avvisa chi di competenza per indagare.

CLASSIFICAZIONE DI IMMAGINI PET
Data un'immagine PET cerebrale, identificare quali sono le zone di interesse (per esempio parti tumorali) ed associare queste informazioni ad una classe.
Se tale procedura viene effettuata per un numero elevato di dati, si può utilizzare il dataset ottenuto per identificare in altre immagini PET la presenza o meno di tali zone di interesse e sapere se un paziente può essere classificato come sano o malato.

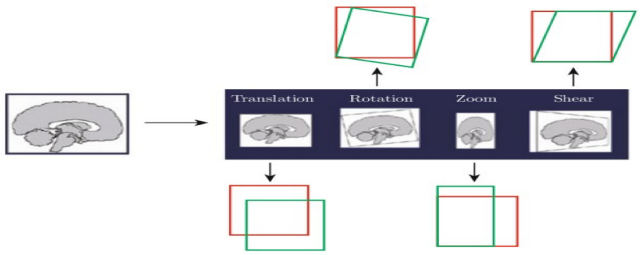


Fig. 13 Normalization step: 12-Parameters Affine Transfor-mations.

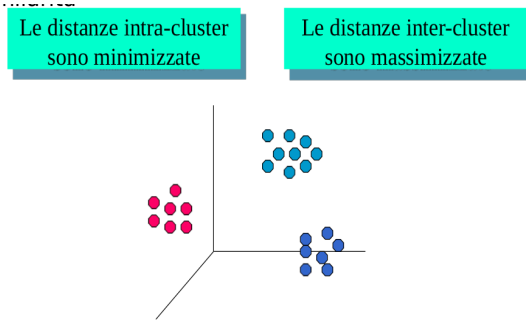
Ricapitolando, arriva la PET di un nuovo paziente, ne estraggo soltanto la porzione di interesse e la si "dà in pasto" al mio modello che mi guida e mi dice se questa immagine può essere associata ad una classe oppure ad un'altra (sano o malato).
È chiaro che se abbiamo utilizzato un dataset per allenare il nostro modello che è fatto da pazienti tra i 40 e 60 anni, qualora si riceva una PET per esempio di un ragazzo di 20 anni, ne risulta un dato che non possiamo utilizzare come target nella classificazione (perché il ragazzo ha una struttura cerebrale con caratteristiche diverse, in termini di immagine PET, rispetto agli individui del dataset che sono più anziani).

CLUSTERING
Fare clustering significa essenzialmente prendere i dati e raggrupparli per similarità.
Definizione: Dato un insieme di oggetti (punti di un iperspazio), ognuno caratterizzato da un insieme di attributi, e avendo a disposizione una misura di similarità tra gli oggetti, trovare i sottoinsiemi di oggetti tali che gli oggetti appartenenti a un sottoinsieme siano più simili tra loro rispetto a quelli appartenenti ad altri cluster.

Misure di similarità
Per raggruppare i dati è necessario individuare le caratteristiche che li accomunano e sapere quanto sono simili tra loro. È importante utilizzare il metodo più adatto per misurare la similarità relativo al tipo di dato in analisi. Un modo per misurare la similarità è la distanza euclidea, applicabile se gli attributi dei punti assumono valori continui. Sono possibili molte altre misure che dipendono dal problema in esame. In ambito biomedicale le similarità possono esser definite rispetto a parametri di interesse medico clinico/comportamentale e similmente le metriche.

Inoltre, è importante decidere quale caratteristica considerare per la misurazione; ad esempio, interessi, sport o attività per fare un clustering per l'iscrizione in dei club.

La stessa situazione si applica all'operazione di clustering nei data set dei pazienti. Per determinare se una nuova persona (x) appartiene a un cluster, dobbiamo verificare se x presenta le caratteristiche di appartenenza a tale gruppo. Non è pratico confrontare x con ognuno degli appartenenti al cluster, quindi prendiamo tutte le caratteristiche, come altezza, bellezza e colore degli occhi, di tutti i dati del nostro cluster e identifichiamo un rappresentante, ad esempio la media. Successivamente, confrontiamo x con il rappresentante e questa fase è nota come **calcolo del centroide** (cioè un oggetto che accomuna tutti i dati da utilizzare in un cluster che viene utilizzato per raffronto con i nuovi individui che si vogliono aggiungere al cluster).



Rappresentazione di un clustering nello spazio 3d costruito utilizzando la distanza euclidea come misura di similarità.

Per **distanza intra-cluster** si intende la differenza di similarità tra elementi dello stesso cluster, che deve essere minimizzato.

La **distanza inter-cluster** è, invece, la differenza di similarità tra elementi di cluster diversi, che deve essere massimizzata per poter distinguere al meglio i diversi cluster.

Un'altra applicazione di clustering può essere la segmentazione del mercato:

- Obiettivo: suddividere i clienti in sottoinsiemi distinti (cluster) da utilizzare come target di specifiche attività di marketing.
- Approccio:
  - ✓ Raccogliere informazioni sui clienti legati allo stile di vita e alla collocazione geografica
  - ✓ Trovare cluster di clienti simili
  - ✓ Misurare la qualità dei cluster verificando se il pattern di acquisto dei clienti appartenenti allo stesso cluster è più simile di quello di clienti appartenenti a cluster distinti

Regole associative

Dato un insieme di record ognuno composto da più elementi appartenenti a una collezione data, produce delle regole di dipendenza che predicono l'occorrenza di uno degli elementi in presenza di occorrenze degli altri. Per esempio, si prende un dataset di scontrini e dalle analisi si vede che c'è una frequenza elevata negli scontrini che associano per esempio pane, coca cola, e latte; questo risultato può essere usato dai supermercati per organizzare gli scaffali in modo tale che queste merci siano vicine per aumentarne le vendite.

TID	Record
1	Pane, Coca Cola, Latte
2	Birra, Pane
3	Birra, Coca Cola, Pannolini, Latte
4	Birra, Pane, Pannolini, Latte
5	Birra, Pannolini, Latte

Regola:
{Latte} --> {Coca Cola}
{Pannolini, Latte} --
>{Birra}

Categoria	# articoli	#correttamente classsificati	%correttamente classsificati
Finanza	555	364	66%
Esteri	341	260	76%
Cronaca nazionale	273	36	13%
Cronaca locale	943	746	79%
Sport	738	573	78%
Intratteniment o	354	278	79%

Altro utilizzo del clustering: dati una serie di articoli (nell'esempio 3204 articoli del Los Angeles Times) raggrupparli per similarità nelle aree tematiche di appartenenza (finanza, esteri, cronaca nazionale, cronaca locale, sport, intrattenimento). Come misura di similarità si utilizza il numero di parole comuni tra i documenti (escluse alcune parole comuni come le congiunzioni, gli articoli, ecc.).