Data Mining e Bioimmagini

ORGANIZZAZIONE E CARATTERISTICHE DEI DATI

Prof. Pierangelo Veltri – 09/10/2023- Autori: Gulizia, Cassalia – Revisionatori: Gulizia, Cassalia

L'organizzazione delle informazioni nella rappresentazione dei dati si rappresenta con una modalità che dispone di un focus organizzativo rispetto ai tipi di attributi in funzione delle proprie caratteristiche (quantitative e qualitative). Si analizzano e si raggruppano le informazioni già in funzione di ciò che serve per avere conoscenza.
Esempi:

- *Si hanno delle bioimmagini – analisi qualitativa*
- *Si hanno delle analisi del sangue – analisi quantitativa*

Le caratteristiche si possono raggruppare in:

- **QUALITATIVE**
  - *Nominali* → i valori degli attributi nominali possono essere *uguali* o *diversi.*
    - Le operazioni che si possono fare sono entropia, correlazione di contingenza e test $\chi 2$
      - Esempi: codici postali, codici identificativi dei dipendenti, colore degli occhi, genere
  - *Ordinali* → i valori degli attributi ordinali possono essere *maggiori* o *minori.*
    - Le operazioni che si possono fare sono mediana, percentili, correlazione di ranghi, test dei run, test dei segni
      - Esempi: durezza dei minerali, gradi, numeri civici
- **QUANTITATIVE**
  - *Intervalli* → per gli attributi di intervallo, le differenze tra i valori sono significative
    - Le operazioni che si possono fare sono media, deviazione standard, correlazione di Pearson, t test e F test
      - Esempi: date del calendario, temperatura in Celsius o Fahrenheit
  - *Mutua relazione* → Per le variabili di rapporto, sia le differenze che i rapporti sono significativi
    - Le operazioni che si possono fare sono media geometrica, media armonica e variazione percentuale
      - Esempi: temperatura in Kelvin, quantità monetarie, conteggi, età, massa, lunghezza, corrente

I dati appartengono sia allo spazio del CONTINUO che del CONCRETO:

- *CONTINUO*
  - I valori reali si possono solo misurare e rappresentare usando un numero finito di cifre. I numeri reali sono i valori degli attributi e questi ultimi sono generalmente rappresentati come variabili a virgola mobile.
    - *Esempi: temperatura, altezza o peso*
- *CONCRETO*
  - Gli attributi concreti hanno solo un insieme di valori finito o numerabile e spesso sono rappresentati come variabili intere. *(gli attributi binari sono un caso speciale di discreti attributi)*
    - *Esempi: codici postali, conti o insieme di parole in una raccolta di documenti*

OPERAZIONI SUI TIPI DI ATTRIBUTI

I tipi di operazioni che si scelgono dovrebbero essere "significativi" per il tipo di dati di cui si dispone.

- **Distinzione, ordine, intervalli significativi e rapporti significativi** sono solo quattro (tra le molte possibili) proprietà dei dati
- Il tipo di dati visualizzato, spesso numeri o stringhe, potrebbe non catturare tutte le proprietà o suggerire proprietà che non sono presenti
- L'analisi può dipendere da altre proprietà dei dati
    - Molte analisi statistiche dipendono solo dalla distribuzione

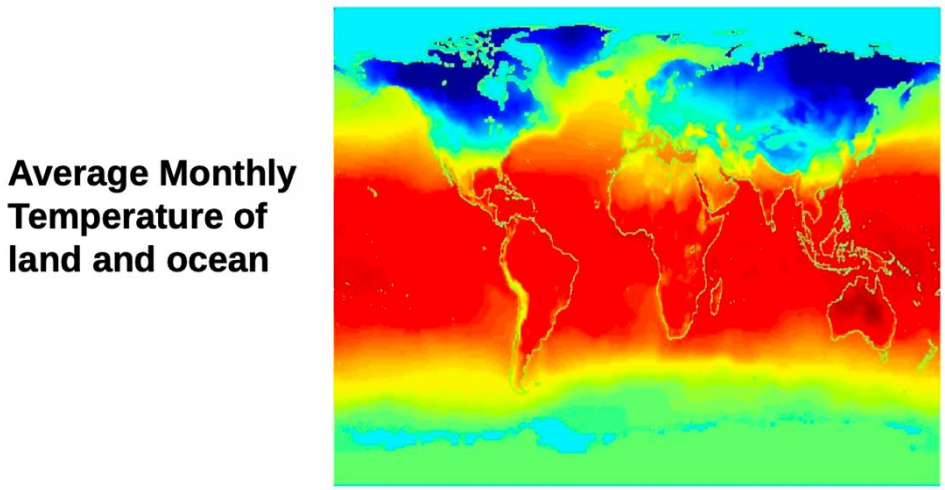
Alla fine, ciò che è significativo può essere specifico del dominio.

I dati possono avere varie caratteristiche:

- Dimensionalità, numeri di attributi → informazioni che si hanno a disposizione (OLAP, online analitical processing) un insieme di caratteristiche descrittive.
    - *Esempio: focus sulle case – informazioni sull'indice energetico, numero di posti letto, metratura, piano, feedback sul quartiere, anche il passaparola è una dimensione di informazioni*
- Scarsità → conta solo la presenza
- Risoluzione → i modelli dipendono dalla scala, è una sorta di affidabilità *(quel dato lo si trova sempre?!).* Si hanno dei missing values, ovvero delle informazioni assenti in tutti i dati oppure non sempre presenti.
- Taglia → il tipo di analisi può dipendere dalla dimensione dei dati

PENTHAO *prende i dati e li analizza utilizzando la lettura delle dimensioni.*
Esempio: Il general manager della casa automobilistica usa un cruscotto direzionale, interfaccia, che consente di leggere gli indicatori per capire come sta andando l'andamento della vendita delle auto.

Un tipo di dati è rappresentato da quelli ***spazio-temporali***, dati che riguardano eventi accaduti e che accadono con un valore temporale che si ripete ***(time-series)***. L'attributo tempo ha un'informazione caratteristica, anche se ripetitiva, da cui si trae fuori un indicatore.
I valori e le informazioi hanno caratteristiche di latitudine e di longitudine.



Esempio: la SARS 2003 rimase confinata perché si è sviluppata in una zona dove il movimento era ridotto, la gente si muoveva poco a piedi, quindi ancora meno con i mezzi.

QUALITÀ DEI DATI

La qualità del dato inteso come *affidabilità dell'informazione* influenza le analisi dei dati stessi. La classificazione dei dati relativi a classi di sani e di malati potrebbe esser fortemente influenzata da dati caricati e ottenuti:

- Su intervalli di tempo non continui e non affidabili
- Su dati che contengono errori di valutazione
- Su dati che contengono valori mancanti

I dati sono influenzati da informazioni quali il **rumore** (noise) che portano a modifiche di dati o di misurazioni e riguardano anche la trasmissione in remoto di dati di controllo ECG e parametri vitali provenienti dalle terapie intensive.

Per migliorare le analisi che si fanno sui dati, il rumore deve essere eliminato e può essere fatto sia in fase di analisi diretta sia nel caso in cui si prendano dei dati e si vogliano analizzare. Si può anche avere la necessità di considerare:

- ***Precisione***: indica la similarità di sequenza di misurazioni. Si misura con deviazione standard;
- ***Bias***: la variazione tra le misurazioni. Si prendono le misurazioni e si fa una media e il bias è la differenza tra il valore reale e quello della media;
- ***Accuratezza***: implica la precisione. Una delle misure, per esempio, è il numero di cifre significative dopo lo zero.

Sono presenti, inoltre, dei valori che sono al di fuori delle misurazioni nel complesso. Questi vengono chiamati **outliers,** che possono essere eliminati perché sono identificati come rumore ma anche valutati come identificatori di valori diversi ovvero valori di interesse.

Nel data mining si hanno tanti dati che arrivano da vari sensori e devono essere valutati tutti.

Tra i dati ci possono essere anche dei **missing values:**

- Valori mancanti
  - Dati non raccolti *(esempio: le persone si dimenticano di dare la propria età)*
  - Attributi non applicabili *(esempio: il reddito annuo non è applicabile ai bambini)*
  - Dimenticanza
- Gestione dei dati mancanti
  - Elimino i dati e i loro correlati
  - Stima dei valori mancanti
  - Ignorare i dati

DATA PREPROCESSING

I passi di preprocessing dei dati servono a risolvere e organizzare i dati in modo da poterli manipolare. Il tutto si basa su dei dati scelti. In particolare, si ha:

- Aggregazione
- Campionamento
- Riduzione della dimensionalità
- Discretizzazione e binarizzazione
- Trasformazione degli attributi
- Selezione del sottoinsieme di funzionalità
- Identificazione delle caratteristiche

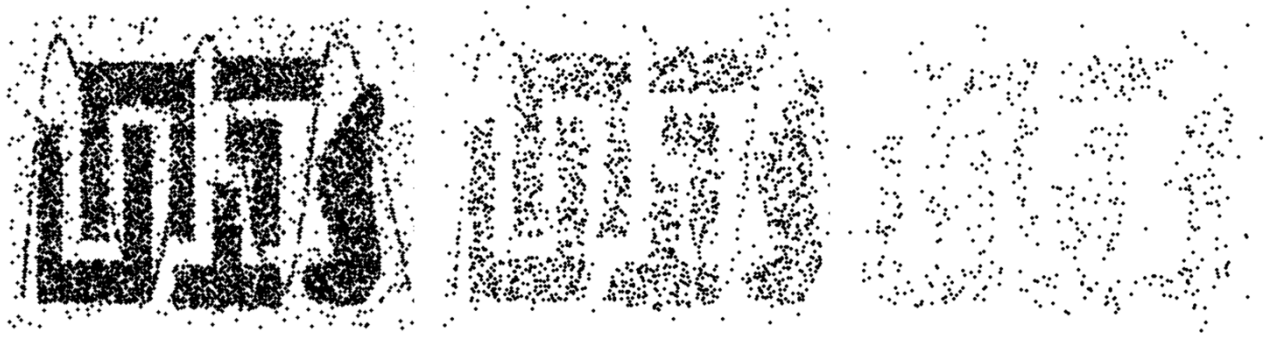
AGGREGAZIONE
Per consentire di utilizzare gli algoritmi di estrazione di dati, sui dati stessi si fa un'aggregazione ovvero si raggruppano dati su una o più informazioni.

Table 2.4. Data set containing information about customer purchases.

Transaction ID	Item	Store Location	Date	Price	...
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	...
101123	Battery	Chicago	09/06/04	$5.99	...
101124	Shoes	Minneapolis	09/06/04	$75.00	...
⋮	⋮	⋮	⋮	⋮	

CAMPIONAMENTO
Il campionamento è la tecnica principale utilizzata per la riduzione dei dati. Serve a estrarre informazioni dal dataset. Viene spesso utilizzato sia per l'indagine preliminare dei dati che per l'analisi finale dei dati. Il campionamento viene generalmente utilizzato nel data mining perché l'elaborazione dell'intero set di dati di interesse è troppo costosa o richiede molto tempo.
Il principio chiave per un campionamento efficace è il seguente:

- l'utilizzo di un campione funzionerà quasi altrettanto bene dell'utilizzo dell'intero set di dati, se il campione è rappresentativo
- un campione è rappresentativo se ha approssimativamente le stesse proprietà (di interesse) dell'insieme di dati originale



8000 points **2000 Points** **500 Points**

RIDUZIONE DELLA DIMENSIONALITÀ
Target:

- Evitare il problema della dimensionalità;
- Ridurre la quantità di tempo e memoria richiesti dagli algoritmi di data mining;
- Consentire una visualizzazione più semplice dei dati;
- Può aiutare a eliminare funzionalità irrilevanti o a ridurre il rumore.

Tecniche:

- Analisi dei componenti principali (PCA)
- Scomposizione di un valore singolo
- Altri: tecniche supervisionate e non lineari

In caso di bioimmagini (DICOM) la tecnica PCA fa sì che si ottenga un'immagine brutta ma, nonostante ciò, conserva la forma e alcune caratteristiche utili a conseguire il lavoro.

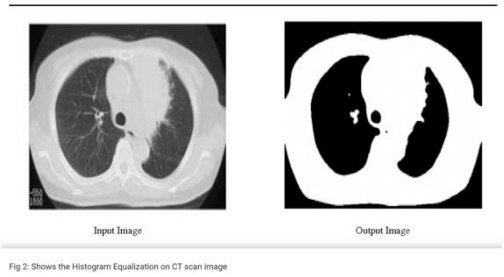


Fig 2: Shows the Histogram Equalization on CT scan image

METODO:
- Pulizia foto
- Estrazione caratteristiche
- PCA – Principals Compontent Analysis

Successivamente vengono effettuati i processi di classificazione-clustering-predizione.

Tra le operazioni più comuni sull'istogramma vi è l'equalizzazione, Questo metodo di solito incrementa il contrasto globale di molte immagini, specialmente quando i dati usabili dell'immagine sono rappresentati da valori di intensità molto vicini. Attraverso questo adattamento, le intensità possono essere meglio distribuite sull'istogramma. Questo permette per le aree a basso contrasto locale di massimizzarlo. L'equalizzazione dell'istogramma si ottiene ciò spalmando la maggior parte dei valori di intensità frequente.

Esempio di R:
Consideriamo il dataset life. Questo dataset contiene i valori di vita attesa nel 1960, anna nascita a 25, 50 e 75 anni di età per uomini e donne in 21 paesi e regioni del mondo.

```r
library(pdataita)
data("life")
str(life)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    31 obs. of  9 variables:
##  $ country: chr  "Algeria" "Cameroon" "Madagascar" "Mauritius" ...
##  $ m0     : num  63 34 38 59 56 62 50 65 56 69 ...
##  $ m25    : num  51 29 30 42 38 44 39 44 46 47 ...
##  $ m50    : num  30 13 17 20 18 24 20 22 24 24 ...
##  $ m75    : num  13 5 7 6 7 7 7 7 11 8 ...
##  $ w0     : num  67 38 38 64 62 69 55 72 63 75 ...
##  $ w25    : num  54 32 34 46 46 50 43 50 54 53 ...
##  $ w50    : num  34 17 20 25 25 28 23 27 33 29 ...
##  $ w75    : num  15 6 7 8 10 14 8 9 19 10 ...
```

In tal caso si iniziano a ridurre delle informazioni, ovvero solo quelle utili al fine delle analisi, oppure si utilizzano delle tecniche automatiche come lo *scatter plot (grafico di dispersione).*

TECNICHE RELATIVE ALLA RIDUZIONE DELLA DIMENSIONALITÀ
Una di queste tecniche si basa sulle caratteristiche
- **Identificare nel dataset un sottoinsieme di attributi**
- Eliminare feature ridondanti e feature irrilevanti
- Esistono tre approcci: embedded, filter, wrapper
    - Embedded: l'algoritmo decide quali selezionare e quali rifiutare (albero decisionale);
    - Filter: selezionare prima dell'esecuzione degli algoritmi;
    - Wrapper: si usano i dati e gli attributi target dell'algoritmo stesso.

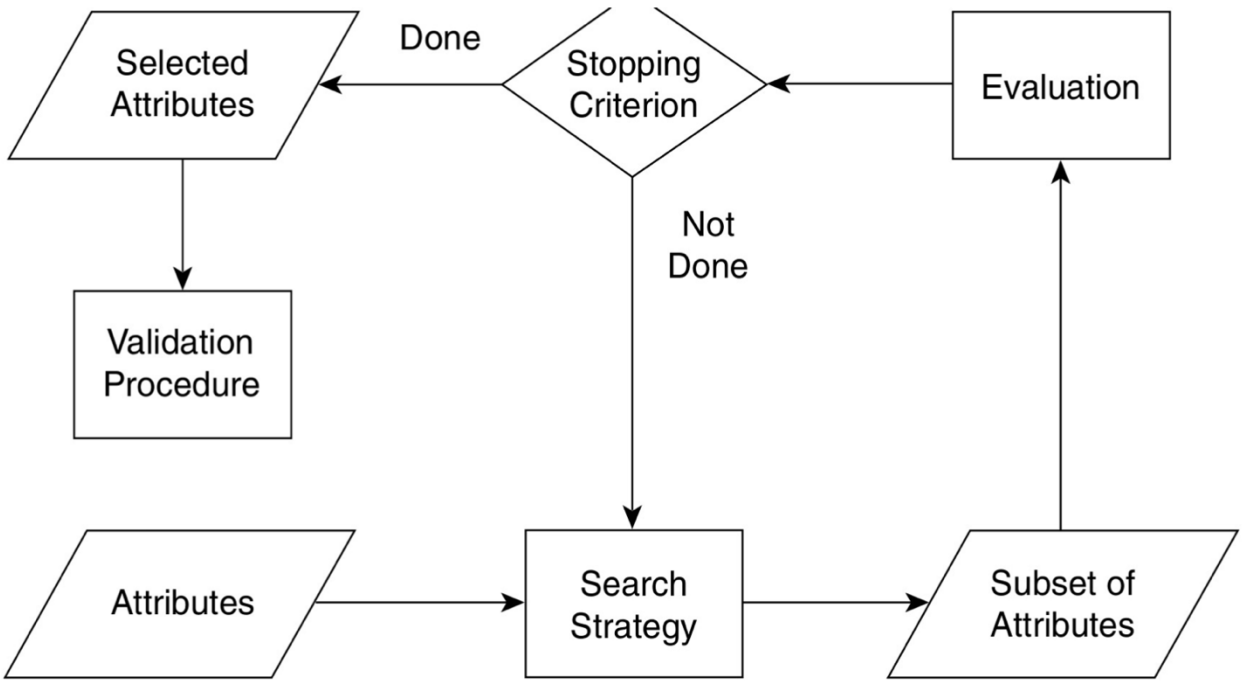


Figure 2.11. Flowchart of a feature subset selection process.

Esistono delle altre tecniche di identificare e togliere i features. È possibile avere:

- **Funzionalità ridondanti**
  - Duplicare gran parte o tutte le informazioni contenute in uno o più altri attributi;
  - *Esempio: prezzo di acquisto di un prodotto e importo dell'imposta sulle vendite pagata.*
- **Caratteristiche irrilevanti**
  - Non contenere informazioni utili per i dati;
  - *Esempio: l'ID degli studenti è spesso irrilevante per il compito di prevedere il GPA degli studenti;*

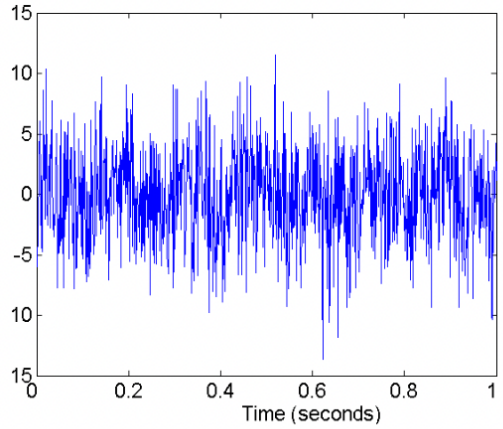
Tra le due features è possibile ritrovarne qualcuna che crea rumore (ovvero dà fastidio). Da un insieme di attributi che descrivono i dati se ne possono estrarre di nuovi.

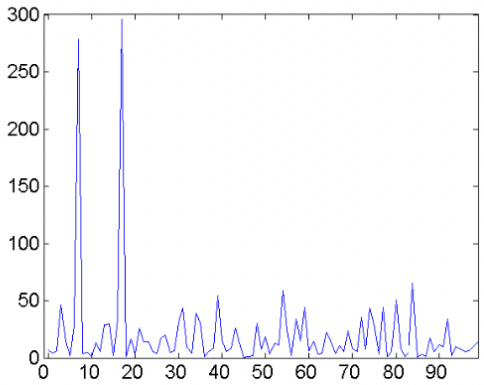
- *Esempio: un insieme di foto (con attributi set of pixel) seleziono solo la combinazione che da immagini di volto*
- *Esempio: da un insieme di valori numeri raggruppati per attributi di descrizione analiti estraggo indicatori per ulteriori analisi (quello che si fa nelle indagini per neoplasie)*

È possibile creare nuovi attributi in grado di acquisire informazioni importanti in un set di dati in modo molto più efficiente rispetto agli attributi originali

- Esistono **tre** metodologie generali:
  - **Estrazione di caratteristiche**
    - Esempio: estrazione dei bordi dalle immagini
  - **Costruzione delle caratteristiche**
    - Esempio: dividere la massa per il volume in modo da ottenere la densità
  - **Mappatura dei dati in un nuovo spazio**
    - Esempio: analisi di Fourier e wavelet

- **Segnale nel tempo allo spazio di Fourier e allo spazio wavelet**





Frequency

La *funzione di mapping elementare* creata nello spazio è una funzione di trasformazione ed è possibile passare da una funzione elementare ad una funzione che si propaga nel tempo e nello spazio.
Esempio: Se volessi generare un'onda senza muovere le mani, viene difficile nello spazio che conosciamo, se non impossibile. Quest'onda in un altro spazio si trasforma in modo diverso.

Alla base del concetto di trasformazione si ha l'utilizzo delle features per generarne delle altre. Quando si fa mapping non si deve avere perdita di informazione e dipende dal target.