Data Mining e Bioimmagini

**DATA MINING E APPLICAZIONI**

Prof. Pierangelo Veltri – 02/10/2023- Autori: Gulizia, Cassalia - Revisionatori: Gulizia, Cassalia

Una parte rilevante dell’ultimo decennio è legata alla disponibilità immediata di raccogliere dati e metterli a disposizione. Tra la fine degli anni 90 e inizi 2000, ci si basava molto sull’estrarre dati e informazioni a partire da pagine web; oggi questo meccanismo viene fatto in maniera online (come, ad esempio, tramite l’utilizzo di Alexa o ChatGPT). La parte interessante è l’acquisizione dei dati in quanto la prima direction non si basa sulla necessità della singola persona, ma la direzione che ha trainato il mondo dell’acquisizione dei dati è legata alla parte delle grandi linee commerciali. Bisogna fornire servizi personalizzati affinché ci sia un vantaggio di tipo competitivo.

*OVUNQUE SI HANNO GRANDI QUANTITÀ DI DATI*

Si verifica un’enorme crescita dei dati in database commerciali e scientifici a causa dei progressi nella generazione di dati e tecnologie di raccolta.

**Nuovo mantra**: raccogli tutti i dati che puoi quando e dove possibile.

**Aspettative**: i dati raccolti avranno valore sia per lo scopo per cui sono stati collezionati o per uno scopo non previsto.

**PUNTO DI VISTA COMMERCIALE**

Molti dati vengono raccolti e immagazzinati

* Dati web
* Google ha peta byte di dati web
* Facebook ha miliardi di utenti attivi
* Acquisti ed e-commerce
* Amazon gestisce milioni di visite al giorno
* Transazione bancarie/ con carta di credito
* I computer sono diventati più economici e più potenti
* La pressione competitiva è forte
* Bisogna fornire servizi personalizzati migliori per un vantaggio competitivo (e.g. Customer Relationship Management)

Immagine che contiene testo, schermata, Carattere, design

Descrizione generata automaticamente

**ChatGPT** è basato su una serie di algoritmi tali che è in grado di rispondere a qualsiasi richiesta. Qui di fianco un’immagine riassuntiva ma molto provocatoria.

**Bard** è stato sviluppato come concorrente diretto di ChatGPT.

Tutta la parte interessante dove sta?

Parlando di **YouTube**, streaming per eccellenza, non esiste un’altra azienda concorrente perché per poter fare una piattaforma alternativa si ha la necessità di convincere l’intera popolazione al fatto che si è migliori, ma se non lo si è la gente continua a navigare su YouTube.

Un esempio è rappresentato da **Google** che ha fatto in modo di avere un grande impatto visivo grazie alla pagina completamente vuota, è presente solo il colore della scritta; si parla quindi di un marker commerciale a d’oc dando un impatto notevole.

**PUNTO DI VISTA SCIENTIFICO**

* Dati raccolti e archiviati sempre più velocemente
* Sensori remoti su un satellite
* Archivi della NASA EOSDIS hanno petabyte di dati di osservazioni della terra per anno
* Telescopi che scansionano i cieli
* Dati del sondaggio Sky
* Dati biologici
* Simulazioni scientifiche
* Terabyte di dati generati in poche ore
* Il data mining aiuta gli scienziati
* Nell’analisi automatizzata di enormi set di dati
* Nella formazione di ipotesi

Oggi la disponibilità di spazio di memoria è elevata, tanto è vero che nel nostro dispositivo personale abbiamo delle applicazioni di intelligenza artificiale. Quando cerchiamo delle immagini possiamo farlo ad esempio per nome, data, oggetto ed è questo quello che fa la differenza.

**COSA È IL DATA MINING?**

L'*estrazione di dati* o *data mining* è l'insieme di tecniche e metodologie che hanno per oggetto l'estrazione di informazioni utili da grandi quantità di dati, attraverso metodi automatici o semi-automatici e l'utilizzo scientifico, aziendale, industriale o operativo delle stesse. L’estrazione può avvenire in diversi modi:

* **Estrazione complessa di informazioni implicite**, precedentemente sconosciute e potenzialmente utili dai dati. Si estraggono dei dati che non sono visibili utilizzando dei metodi statistici
  + *esempio*: *dall’analisi dei dati si dice che la vendita degli ovetti Kinder aumenta se si mette vicino alle creme di bellezza. Non è un’informazione che la statistica tira fuori, ma è un’informazione che il data mining con qualche tool può ricavare. Esiste una potenziale associazione nei dati relativi alle vendite tra chi compra una crema e chi compra gli ovetti Kinder.*
* Immagine che contiene diagramma, schermata, testo, Carattere

  Descrizione generata automaticamente**Esplorazione e analisi**, per mezzo di sistemi automatici e semi-automatici, di grandi quantità di dati al fine di scoprire pattern significativi

**Immagine che contiene testo, Carattere, linea, diagramma

Descrizione generata automaticamente**

*Data Preprocessing* 🡪 se si vuole far valutare qualcosa da un algoritmo, bisogna cercare di minimizzare le cose che possono dar fastidio all’analisi; ad esempio, se si mandano tanti report da analizzare e in qualcuno mancano i dati, il risultato è poco significativo.

*Data Postprocessing* 🡪 si considerano i risultati e questi ultimi devono essere resi visualizzabili; ad esempio, prendendo un foglio excel con 100 colonne aventi i risultati, essi devono essere interpretati in modo tale da velocizzare l’analisi che deve essere fatta.

***Da questo processo si otterranno dei risultati che vengono forniti all’esperto di dominio che fa un esame di conoscenza.***

**PATTERN (O MODELLO)**

Un modello è una semplificazione di una parte di realtà che consente di regolamentare le relazioni che esistono tra le variabili o tra gli eventi all’interno di uno spazio chiuso.

Il primo target da fare in data mining è identificare un pattern, ovvero una rappresentazione sintetica e ricca di semantica di un insieme di dati, in grado di poter dare delle informazioni di correlazione tra gli oggetti o tra i dati stessi. Esprime in genere un modello ricorrente nei dati, ma può anche esprimere un modello eccezionale.

﻿﻿Un pattern deve essere:

* **Valido** sui dati con un certo grado di confidenza
* **Comprensibile** dal punto di vista sintattico e semantico, affinché l'utente lo possa interpretare
* Precedentemente **sconosciuto** e potenzialmente **utile**, affinché l'utente possa intraprendere azioni di conseguenza

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamente

*esempio: si hanno delle coppie di valori prestito-stipendio e si stabilisce un modello, ovvero se c’è una soglia al di sopra della quale i pagamenti sono affidabili e al di sotto non lo sono.*

***Tipi di pattern(modelli)***

* **Regole associative**

consentono di determinare le *regole di implicazione logica* presenti nella base di dati, quindi di individuare i gruppi di *affinità tra oggetti* (ad esempio: se si ha il colesterolo alto, il rischio di problemi cardiaci è più elevato rispetto a chi ha dei valori normali)

* **Classificatori**

consentono di derivare un modello per la *classificazione di dati* secondo un insieme di classi assegnate a priori (ad esempio: si analizza un’intera popolazione, si classificano i pazienti sani e i pazienti malati, si prende un nuovo dato e lo si associa a sano o malato)

* **Alberi decisionali**

sono particolari classificatori che permettono di identificare, in ordine di importanza, le cause che portano al verificarsi di un evento (ad esempio: arriva un paziente, si controllano i dati clinici e si effettua una risonanza magnetica. Se risulta positiva o negativa, si considera una “strada” diagnostica fino ad arrivare all’identificazione a quello che è lo status del paziente in quel momento stesso)

* **Clustering**

raggruppa gli elementi di un insieme, a seconda delle loro caratteristiche, in classi non assegnate a priori

* **Serie temporali**

permettono l'individuazione di pattern ricorrenti o atipici in sequenze di dati complesse

Immagine che contiene testo, schermata, Carattere, diagramma

Descrizione generata automaticamente

**RAPPRESENTAZIONE DEI DATI**

I dati devono essere messi in una propria modalità di rappresentazione.

*Esempi*

1. Pazienti sani e malati: dato un insieme di pazienti descritti da features (ghiandola prostatica, età, proteine...) e da una classe (“sano”, “malato”) identificare: (I) appartenenza a una delle due classi guardando soltanto una combinazione di features
2. 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 (classificazione)
3. Dato un insieme di segnali vocali associati a pazienti malati di SLA e pazienti neurologici, identificare la patologia o la tendenza a sviluppare
4. Brain aging: dato un insieme di immagini RM cerebrale identificare l’appartenenza ad una classe di età usando: immagini e patologie
5. Dataset Iris (fiori) date le misure dei petali e dello stelo, identificare l’appartenenza di un fiore ad una classe (famiglia)

**ATTIVITÀ TIPICHE DEL DATA MINING**

* Sistemi di **predizione**

Utilizzare alcune variabili per predire il valore incognito o futuro di altre variabili.

* Sistemi di **descrizione**

Trovare pattern interpretabili dall’uomo che descrivano i dati

1. **Classificazione** [Predittiva] si hanno le classi e voglio spiegarle e descriverli in termini di caratteristiche
2. **Clustering** [Descrittiva] – scoprire gruppi di dati simili
3. Ricerca di **regole associative** [Descrittiva]
4. Ricerca di **pattern sequenziali** [Descrittiva]
5. **Regressione** [Predittiva] descrivere le forme (funzione) che approssima i dati
6. Individuazione di **deviazioni** [Predittiva]