Data Analytics

Giuseppe Magazzù

2021 - 2022

Contents

Chapter 1

Introduzione

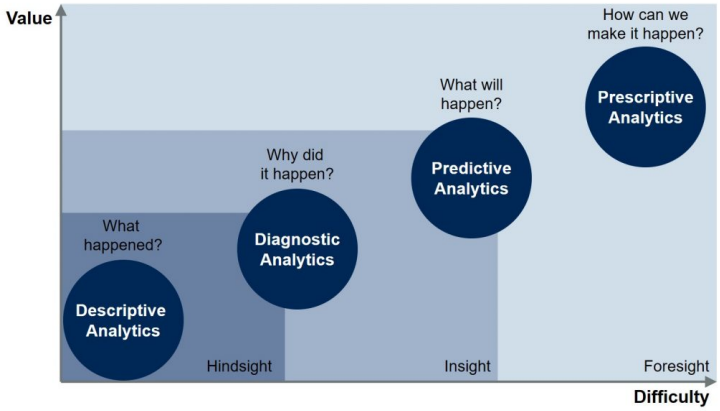

Figure 1.1: Diversi tipi di analisi per valore e difficoltà [3].

1.1 Definizioni

Un'**istanza** (instance, item, record) é un esempio descritto da un insieme finito di attributi. Il numero di attributi può variare per alcune istanze.

Un **attributo** (attribute, field, variable) é una misura di un aspetto di un'istanza.

Tipi di attributi:

- **Quantità nominali**: i valori sono **simboli** distinti. Non hanno relazioni come ordinamento o distanza.
  (e.g. attributo: "outlook", valori: "sunny", "cloudy", and "rainy").
- **Quantità ordinali**: i valori hanno una relazione d'ordine, ma non di distanza.
  (e.g. attributo: "temperature", valori: "hot" > "mild" > "cold").
- **Quantità d'intervallo**:
- **Quantità di rapporto**:

Una **classe** (class, label) rappresenta un gruppo di istanze che condividono delle caratteristiche comuni.

Propositionalization

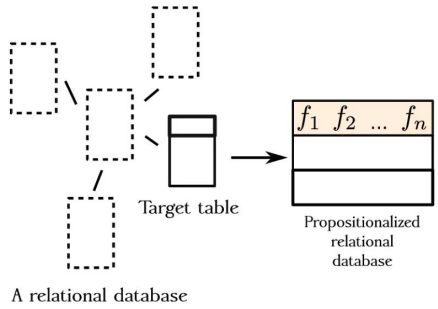

Figure 1.2: Processo di propositionalization [2]

1.2 Pre-Processing

I dati nel mondo reale sono **incompleti**, **rumorosi** e **inconsistenti**. Per ottenere dell'analisi di qualità é necessario effettuare prima delle operazioni sui dati.

- **Data Cleaning**: sostituire valori mancanti, smussare dati rumorosi, identificare o rimuovere outliers e risolvere inconsistenze.
- **Data Integration**: integrazione di diversi dati.
- **Data Transformation**: normalizzare o aggregare i dati
- **Data Reduction**: feature selection, feature extraction

1.3 Data Cleaning

Dati Mancanti

Alcuni dati possono non essere stati calcolati o possono non essere disponibili per malfunzionamenti o per errori umani.

L'assenza di questi dati **complica l'analisi** poiché non tutti i metodi di analisi non gestiscono questo problema, inoltre comporta una **perdita di efficacia** nell'estrarre dei pattern.

Categorie di valori mancanti:

- **Missing Completely At Random (MCAR)**:
- **Missing At Random (MAR)**:
- **Not Missing At Random (NMAR)**:

Gestione dei valori mancanti:

- **Ignorare** le istanze o gli attributi con valori mancanti. Praticabile solo se ci sono pochi esempi mancanti poiché introdurrebbe un bias.
- **Convertire** i valori mancanti in un nuovo valore ("missing", "?", "NA").
- **Imputare** i valori mancanti basandosi sul resto del dataset.

Metodi di Imputazione

- **Most Common (MC) Value**
  Assunzione: ogni attributo ha una distribuzione normale.
  - Valori **continui**: rimpiazza con la media dell'attributo nel dataset
  - Valori **discreti**: rimpiazza con il valore più frequente dell'attributo nel dataset

- **Concept Most Common (CMC) Value**
  Assunzione: ogni attributo ha una distribuzione normale per tutte le istanze che appartengono alla stessa classe.

  I valori mancanti vengono rimpiazzati con il valore medio/più frequente delle istanze della stessa **classe**.

- **K-Nearest Neighbors**
  Le istanze vengono disposte in uno spazio metrico e i valori mancanti vengono imputati considerando le k istanze più vicine.

Dati Rumorosi

Alcuni dati possono avere errori dovuti a **strumenti difettosi**, **errori umani** o **di calcolo**, errori durante la **trasmissione** dei dati o **limitazioni tecnologiche**.

Questi errori introducono del "rumore" all'interno dei dati che può essere rimosso usando tecniche di **data smoothing**. Queste tecniche riducono il rumore e rendono i pattern più identificabili, tuttavia si riduce la quantità di dati da analizzare e inoltre gli outliers possono alterare l'analisi.

Binning

I dati vengono **ordinati** e **partizionati** in bin. Quindi ogni bin si può smussare con media, mediana dei valori all'interno o utilizzando gli estremi.

- **Equal-width** (distance) partitioning: viene diviso il range in N intervalli di uguale dimensione.
- **Equal-depth** (frequency) partitioning: viene diviso il range in N intervalli, ognuno dei quali contiene approssimatamene lo stesso numero di esempi.

Dati Sbilanciati

Esistono molti problemi di classificazione in cui una classe ha una distribuzione fortemente sbilanciata, ovvero che il numero di osservazioni per una classe è molto inferiore a un'altra (e.g. fraud detection, disease diagnosis, natural disaster, etc.). Quindi risulta difficile ottenere buoni valori di accuracy su entrambe le classi.

Un possibile approccio è quello di bilanciare i dati del train set.

- **Oversampling**: aggiungere istanze alla **classe minoritaria** tramite campionamento con rimpiazzo (i.e. duplicare alcuni valori) fino a ottenere lo stesso numero di istanze per classe. Bilancia le classi, ma non fornisce nuove informazioni al modello.
- **Undersampling**: rimuovere randomicamente istanze dalla **classe maggioritaria** fino a ottenere lo stesso numero di istanze per classe.

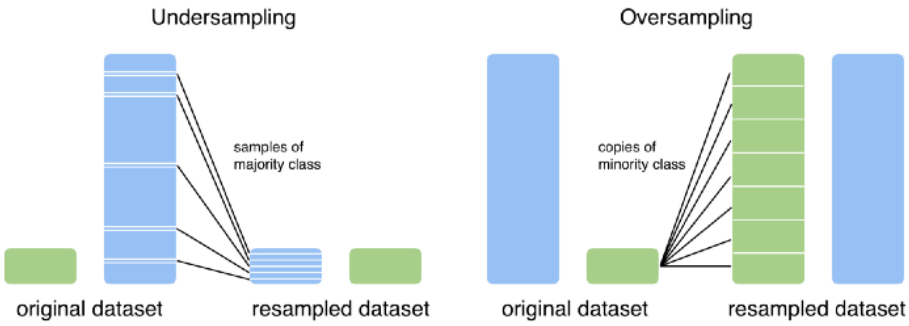

Figure 1.3: Rappresentazione del funzionamento del random resampling.

Synthetic Minority Oversampling Technique (SMOTE)

SMOTE è una tecnica di oversampling che genera esempi sintetici della classe minoritaria a partire dai dati esistenti. Dato un esempio della classe minoritaria vengono selezionati i k esempi più vicini, viene scelto uno a caso tra questi e viene generato un numero esempio tra questi due.

Tomek Links

Tomek Links è una tecnica di undersampling che rimuove gli esempi della classe maggioritaria che appartengono a un Tomek Link. Un **Tomek Link** é una coppia d'istanze di classi diverse per cui non esiste nessun'altra istanza che sia più vicina a uno dei due.

La collezione di Tomek Links nel dataset definisce le frontiere delle classi.

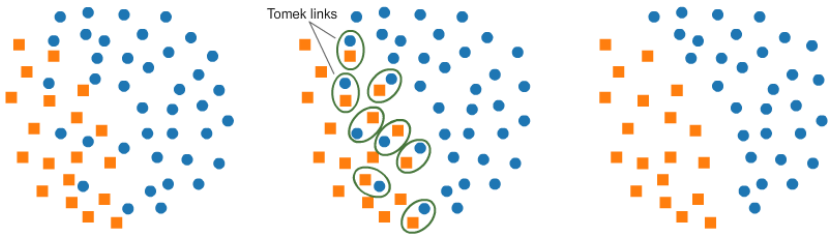

Figure 1.4: Esempio di undersampling tramite Tomek Links [1].

Chapter 2

Networks Analytics

2.1 Statistiche Descrittive

Grado

Il **grado** k di un nodo i è il numero di archi entranti e uscenti. E' possibile distinguere il grado in **outdegree** k_i^{out} e **indegree** k_i^{in}, il grado totale è dato dalla somma $k = k_i^{\text{in}} + k_i^{\text{out}}$.

In un grafo diretto si definisce **source** un nodo con $k_i^{\text{in}} = 0$, e **sink** $k_i^{\text{out}} = 0$.

Grado Medio

Grafi non diretto:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \qquad\qquad \langle k \rangle = \frac{2|E|}{N}$$

Grafi diretto:

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} \qquad\qquad \langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out}$$

$$|E| = \langle k^{in} \rangle = \langle k^{out} \rangle \qquad\qquad \langle k \rangle = \frac{|E|}{N}$$

Distribuzione del Grado

Si definisce $P(k)$ la probabilità di un nodo di avere grado k e rappresenta la **distribuzione del grado** $P(k) = \mathrm{N}_k/\mathrm{N}$, dove N_k é il numero di nodi con grado k.

Distanza e Diametro

La **distanza** tra due nodi A e B è definita come il numero di archi lungo lo *shortest path* da A a B. Se i due nodi non sono **raggiungibili** tra loro la distanza é infinito.

In un grafo diretto la distanza tra due nodi non é simmetrica, $d(A, B) \neq d(B, A)$.

Per calcolare le distanze in un grafo si utilizza la **breadth-first search (BFS)**.

Il **diametro** di un grafo é la distanza massima tra ogni coppia di nodi nel grafo.

La **distanza media** in un grafo si definisce:

$$\langle d \rangle = \frac{1}{2|E_{max}|} \sum_{i,j \neq i} d_{ij}$$

grafo diretto

$$\langle d \rangle = \frac{1}{|E_{max}|} \sum_{i,j > i} d_{ij}$$

grafo indiretto

Coefficiente di Clustering

Il **coefficiente di clustering** é una misura di quanto i nodi vicini tendono a collegarsi tra loro. Dato un nodo i con grado k_i il coefficiente di clustering é definito come:

$$C_i = \frac{2|E_i|}{k_i(k_i - 1)}$$

grafo diretto

$$C_i = \frac{|E_i|}{k_i(k_i - 1)}$$

grafo indiretto

in cui l'insieme $E_i = \{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}$ contiene tutti gli archi che sono collegati al nodo i e $N_i = \{v_j : e_{ij} \in E \vee e_{ji} \in E\}$ é il suo **vicinato**, ovvero i nodi collegati direttamente a i.

Il coefficiente di clustering assume un valore tra 0 e 1:

- Se $C_i = 0$, allora nessun nodo del vicinato sarà collegato agli altri.
- Se $C_i = 1$, allora tutti nodi del vicinato saranno collegati tra loro.

E' possibile catturare il grado di clustering di un grafo con il **coefficiente medio di clustering**, ovvero la probabilità che dato un nodo qualsiasi due nodi nel suo vicinato siano collegati tra loro.

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$

Centralizzazione

La **centralizzazione** di una rete é un indicatore di quanto importate (centrale) sia un nodo rispetto agli altri rispetto a una certa misura di centralità.

Sia $C_x(i)$ una misura di centralità per un nodo i e $C_x(n^*)$ il valore massimo di centralità. Si definisce centralizzazione della rete per questa misura di centralità:

$$C_x = \frac{\sum_{i=1}^{N} C_x(n^*) - C_x(i)}{\max \sum_{i=1}^{N} C_x(n^*) - C_x(i)}$$

Degree Centrality

La misura di **grado** quantifica il numero di archi incidenti a un nodo, ovvero il numero di nodi che lo raggiungono direttamente.

$$C_D(i) = k_i^{\text{in}}$$

Centralizzazione:

$$C_D = \frac{\sum_{i=1}^{N} C_D(n^*) - C_D(i)}{(N-1)(N-2)}$$

Betweenness Centrality

La misura di **betweenness** quantifica il numero di volte che un nodo si comporta da "ponte" in uno *shortest path* tra coppie di nodi.

La betweenness del nodo i si definisce:

$$C_B(i) = \frac{\sum_{j \neq k} g_{jk}(i)}{g_{jk}}$$

dove g_{jk} é il numero di *shortest path* tra i nodi j e k, e $g_{jk}(i)$ é il numero di questi percorsi che passano per il nodo i.

E' possibile normalizzare questa misura dividendo per il numero di coppie di vertici che escludono il nodo considerato.

$$C'_B(i) = C_B(i)/[(N-1)(N-2)/2]$$

Closeness Centrality

La misura di **closeness** quantifica la distanza media di uno *shortest path* da un nodo a tutti gli altri.

$$C_C(i) = \frac{1}{\sum_{j=1}^{N} d(i,j)}$$

dove $d(i,j)$ é la distanza tra il nodo i e il nodo j.

E' possibile normalizzare questa misura moltiplicando per il numero di tutti i vertici del grafo escluso il nodo considerato.

$$C'_C(i) = \frac{N-1}{\sum_{j=1}^{N} d(i,j)} = (N-1)\, C_C(i)$$

Se un nodo ha un valore alto di closeness allora sarà vicino a molti nodi nel grafo. Può essere interpretata come una misura di velocità per raggiungere un qualsiasi nodo della rete.

Eigenvector Centrality

La **eigenvector centrality** misura l'influenza di un nodo all'interno del grafo.

Un nodo é considerato importante se é collegato ad altri nodi importanti. Quindi un nodo che ha tanti archi entranti non ha necessariamente un alto valore di eigenvector centrality.

$$\lambda C_E = A C_E \qquad\qquad (A - \lambda I) C_E = 0, \quad \lambda \neq 0$$

Il valore che assume questa misura di centralità corrisponde all'autovettore C_E associato all'autovalore più grande.

Reciprocity

Dato un grafo diretto, si definisce **reciprocità** il rapporto tra il numero di relazioni ricambiate e il totale di relazioni del grafo.

Due nodi hanno una relazione se esiste almeno un arco che li collega. Una relazione si dice **ricambiata** se esiste un arco in entrambe le direzioni.

Density

La **densità** di un grafo indica quanto il grafo é connesso ed é definita dal rapporto tra il numero di archi della rete e il numero totale di archi possibili.

$$D = \frac{2|E|}{N(N-1)} \qquad\qquad D = \frac{|E|}{N(N-1)}$$

grafo diretto *grafo indiretto*

Un grafo con densità 1 é completamente connesso e prende il nome di **clique**.

2.2 Community Detection

Bibliography

[1] Rahul Agarwal. *The 5 most useful Techniques to Handle Imbalanced datasets*. [Online; accessed 14/03/2022]. 2020. URL: https://mlwhiz.com/images/imbal/1_hubf0730b098fff787d09b5f9aa956817e_24275_500x0_resize_box_2.png.

[2] Nada Lavrač, Blaž Škrlj, and Marko Robnik-Šikonja. "Propositionalization and embeddings: two sides of the same coin". In: *Machine Learning* 109.7 (2020), pp. 1465–1507.

[3] Jason McNellis. *Gartner Four Analytic Types*. [Online; accessed 13/03/2022]. 2019. URL: https://blogs.gartner.com/jason-mcnellis/files/2019/11/GartnerFourAnalyticTypesV5.jpg.