Distanza e somiglianze nella chemiometria compilato finale

Chemiometria (Università degli Studi di Milano-Bicocca)

Scansiona per aprire su Studocu

1. Dettagli del Contributore

Prof. Roberto Todeschini
Gruppo di Ricerca in Chemometria e QSAR
Dipartimento di Scienze della Terra e Ambientali, Università di Milano-Bicocca P.zza della Scienza, 1 – 20126 Milano (Italia)
Telefono: +39 02 64482820
Email: roberto.todeschini@unimib.it

Prof. Davide Ballabio
Gruppo di Ricerca in Chemometria e QSAR
Dipartimento di Scienze della Terra e Ambientali, Università di Milano-Bicocca P.zza della Scienza, 1 – 20126 Milano (Italia)
Telefono: +39 02 64482820
Email: davide.ballabio@unimib.it

Dott.ssa Viviana Consonni (PhD)
Gruppo di Ricerca in Chemometria e QSAR
Dipartimento di Scienze della Terra e Ambientali, Università di Milano-Bicocca P.zza della Scienza, 1 – 20126 Milano (Italia)
Telefono: +39 02 64482820
Email: viviana.consonni@unimib.it

2. Titolo dell'articolo

Distanze e misure di similarità in chemometria e chemoinformatica

3. Nome dei Collaboratori

Roberto Todeschini, Davide Ballabio e Viviana Consonni

4. Affiliazione dei Collaboratori

5. Abstract

Università di Milano-Bicocca, Dipartimento di Scienze della Terra e Ambientali, Piazza della Scienza, 1, Milano, 2014, Italia.

Vengono presentate e discusse diverse misure di similarità/diversità per il data mining, la chemometria e la chemoinformatica in relazione ai diversi dati a cui sono applicate. Dopo una breve presentazione degli assiomi per le funzioni di dissimilarità e similarità, le loro relazioni e il trattamento preliminare dei dati richiesto, vengono fornite le definizioni teoriche e le formule delle misure di distanza e similarità per dati a valori reali, binari, ordinati, di frequenza e di tipo misto, insieme ai concetti principali sulle distanze tra insiemi e meta-distanze. Vengono forniti semplici esempi di calcolo e vengono effettuati confronti estesi sulle distanze definite per dati a valori reali e binari.

6. Parole chiave

chemo-metria
chemo-informatica
misure di
distanza misure
di somiglianza
data mining

assi delle
distanze
meta-distanze
coefficienti binari

7. Testo principale

1 Introduzione

Si può facilmente supporre che i concetti di somiglianza e dissimilarità tra oggetti, eventi, situazioni, ecc. siano sempre stati concetti fondamentali del ragionamento umano, che si basa fortemente sul concetto di analogia.

Le prime tracce esplicite della parola distanza si possono trovare in Aristotele (384 a.C. – 322 a.C.), che, nella sua Metafisica, usava la parola distanza per significare: "È tra le estremità che la distanza è maggiore" o "cose che hanno qualcosa tra di loro, cioè una certa distanza". La distanza ha anche il senso di "dimensione" (come in "lo spazio ha tre dimensioni: lunghezza, larghezza e profondità" (Aristotele, Fisica).

Aristarco di Samo (310 a.C. – 230 a.C.) è stato un astronomo e matematico che per primo stimò la distanza tra la Terra e sia la Luna che il Sole, con una buona approssimazione, nell'unico lavoro conosciuto che gli è solitamente attribuito, Sulle dimensioni e

Distanze del Sole e della Luna.
centre and a distance". The word used in this axiom - διαστηματι

Euclide, uno dei matematici più importanti della storia antica (323 a.C. – 286 a.C.), usò la parola distanza solo nel suo terzo postulato dei Principia: "Ogni cerchio può essere descritto da un

- ha ancora un significato molto generale.

Eratostene di Cirene (276 a.C. – 194 a.C.), matematico, astronomo, geografo e poeta, fu il primo a calcolare la circonferenza della Terra sfruttando la sua stima della distanza tra Alessandria (Egitto) e Siene (l'attuale Aswan, Egitto).

La matematizzazione dei concetti di dissimilarità, diversità, distanza e dei loro termini duali come somiglianza e vicinanza risale allo sviluppo della matematica del XX secolo.

La distanza che utilizziamo nella vita quotidiana è euclidea applicata a spazi 2D o 3D, ma esistono diverse misure di distanza. Ogni distanza ha le proprie caratteristiche, vantaggi e svantaggi. Le distanze sono

utilizzate per misurare la somiglianza tra oggetti rappresentati da un numero esteso di parametri, che è la situazione abituale nella chimica analitica dove gli oggetti sono caratterizzati da diversi segnali, parametri, ecc.

La distanza, a volte chiamata lontananza, è una descrizione numerica di quanto siano distanti tra loro le entità. Nella data mining, le entità sono comunemente oggetti o variabili. Il concetto di distanza è un modo concreto per descrivere cosa significa che gli elementi di uno spazio siano "vicini" o "lontani" l'uno dall'altro.

Il valore numerico di una misura di somiglianza/diversità dipende da tre componenti principali: a) la descrizione degli oggetti (cioè le variabili selezionate), b) lo schema di pesatura degli elementi di descrizione e c) la misura di distanza o somiglianza selezionata.

Le misure di distanza e somiglianza svolgono un ruolo fondamentale nella chemometria, come mostrato schematicamente nella Figura 1, dove i diversi metodi sono arbitrariamente divisi in a) metodi che calcolano una distanza tra entità come oggetti o variabili, b) metodi che calcolano una distanza tra un oggetto e un punto di riferimento, e c) metodi che calcolano una distanza tra due insiemi di oggetti (o variabili).

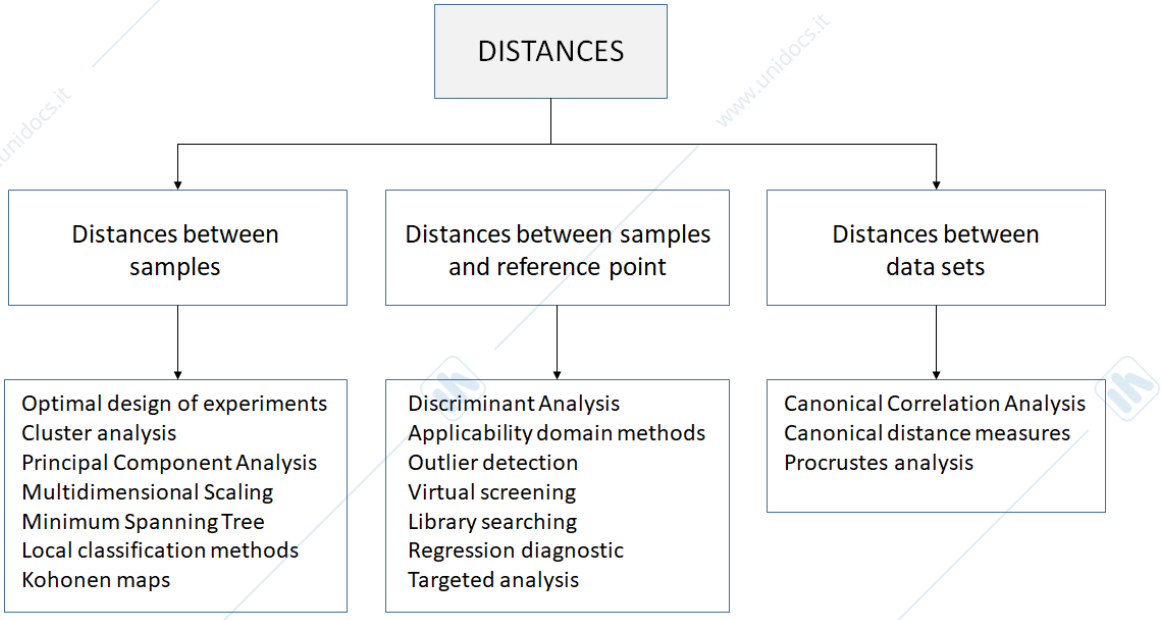

Figura 1 Diagramma di flusso delle diverse applicazioni delle misure di distanza nella data mining e nella modellazione.

Dato i dati, il problema principale nell'utilizzare diversi di questi metodi è la scelta di una distanza appropriata, scelta complicata dal numero enorme di possibilità diverse. Infatti, spesso, gli utenti chemometrici non sono consapevoli delle diverse possibilità che l'uso di misure di distanza/similarità alternative può fornire per evidenziare nuove fonti di informazione.

Le misure di similarità/diversità sono il nucleo di quasi tutti i metodi per l'analisi dei cluster: il metodo k-means assegna un oggetto misurando la sua distanza da ciascun centroide del cluster, che è considerato come il punto rappresentativo del cluster; analogamente, il metodo Jarvis-Patrick si basa su una tabella dei vicini in cui per ogni oggetto da assegnare vengono elencati i vicini più prossimi; i metodi gerarchici agglomerativi utilizzano le misure di distanza, chiamate metriche di collegamento, che sono in grado di quantificare la similarità tra gruppi di oggetti (cioè, i cluster).Le mappe di Kohonen (o Mappe Auto-Organizzanti, SOM) sfruttano le misure di distanza per assegnare oggetti ai neuroni della mappa e poi per valutare le relazioni tra gli oggetti sulla proiezione della mappa: ogni oggetto, rappresentato da un vettore p-dimensionale (dove p è il numero di variabili) scalato nell'intervallo [0, 1], viene confrontato con ciascun neurone della mappa, i cui pesi sono rappresentati nella stessa scala e viene assegnato al neurone per il quale la distanza euclidea è minima (neurone vincitore). Poi, il processo di apprendimento procede e le informazioni ricevute nel neurone vincitore dall'oggetto vengono diffuse ai vicini di quel neurone, smussando le informazioni proporzionalmente alla distanza topologica dal neurone vincitore.

L'Analisi delle Componenti Principali (PCA), che è l'approccio più comune per l'analisi esplorativa dei dati, genera uno spazio metrico in cui le distanze tra le coppie di oggetti sono le classiche distanze euclidee. Altre tecniche comuni per l'analisi esplorativa dei dati, come il Scaling Multi-Dimensionale (MDS) e l'Albero di Copertura Minima (MST), si basano su algoritmi in grado di elaborare matrici di distanza (o similarità) che rappresentano le relazioni di similarità/diversità interne degli oggetti di un insieme di dati [1].

Nella classificazione, uno dei metodi più noti per problemi non lineari è il metodo k-Nearest Neighbour (k-NN), che è tra i metodi di classificazione locale [2] e si basa sul

calcolo di alcune misure di distanza tra l'oggetto target e gli oggetti di addestramento per identificare i primi k vicini e valutare l'appartenenza alla classe dell'obiettivo. Un altro metodo ben noto è l'Analisi Discriminante (DA) che sfrutta la distanza di Mahalanobis, in grado di tenere conto dell'intera struttura di covarianza di ciascuna classe, per valutare la distanza tra l'oggetto e il centroide della classe [1].

Il concetto di dominio di applicabilità del modello (AD) sta guadagnando sempre più importanza nel campo della modellazione, dove la valutazione se un dato modello (modello di classificazione o di regressione) sia adatto o meno a fornire previsioni affidabili per nuovi oggetti è fondamentale [3]. Sono stati proposti diversi metodi per decidere se un nuovo oggetto possa essere considerato all'interno del dominio di applicabilità di un modello e la maggior parte di essi si basa su soglie di distanza predefinite.

Alcuni libri, recensioni e articoli dedicati a presentare, analizzare e confrontare misure di somiglianza/diversità sono [4, 5,6,7,8,9,10,11,12,13].

In questo lavoro, il problema della valutazione delle relazioni di somiglianza e dissimilarità tra oggetti è stato affrontato con particolare attenzione all'uso di misure di distanza e somiglianza nel data mining. Le misure di somiglianza/diversità comunemente utilizzate in altri campi scientifici, come ad esempio quelle basate sul contenuto informativo, non sono state considerate in questo studio. Inoltre, l'articolo si è principalmente concentrato sulle misure di somiglianza/diversità tra oggetti, sebbene alcune misure, come le misure di correlazione di Pearson, Spearman e Kendall, siano solitamente applicate per valutare le relazioni tra variabili.

Nel contesto del data mining, le misure di distanza e somiglianza sono comunemente distinte in base ai tipi di dati, cioè dati a valore reale, dati binari, dati di frequenza o dati classificati, e dati di tipo misto. Pertanto, dopo una breve introduzione al background teorico e al trattamento dei dati richiesto prima del calcolo delle distanze, vengono presentate le misure di distanza/somiglianza più comuni secondo questa classificazione generale, che tiene conto del tipo di dati. Nella parte finale dell'articolo, vengono presentate alcune applicazioni delle diverse misure di distanza e somiglianza a dati reali e simulati.

i set di dati vengono discussi con l'obiettivo di valutare e confrontare le diverse informazioni fornite da ciascuna misura.

2 Contesto teorico

2.1 Notazione e simboli

Un dataset chimico è solitamente costituito da un numero di oggetti (esperimenti, campioni, molecole, ecc.) e da un numero di parametri (variabili, attributi, ecc.), che sono stati misurati su ciascun oggetto. Pertanto, il dataset è organizzato in una matrice numerica (array bidimensionale): ogni riga rappresenta un oggetto del dataset, mentre le colonne rappresentano i parametri chimici. Le matrici numeriche sono denotate come X (n × p), dove n è il numero di oggetti e p il numero di variabili. L'elemento (i − j) della matrice dati X è denotato come xij e rappresenta il valore della j-esima variabile per il i-esimo oggetto. Gli scalari sono indicati da caratteri minuscoli in corsivo (ad es. xij) e i vettori da caratteri minuscoli in grassetto (ad es. x).

2.2 Regole assiomatiche per le misure di dissimilarità

Una funzione $D: \quad X \times X \to \mathbb{R}$, dove X è un insieme, deve soddisfare un certo numero di proprietà (assiomi) per essere considerata una distanza.

Sia X un insieme. Una funzione $D: \quad X \times X \to \mathbb{R}$ is called a *distance* on X if, for all $x, y \in X$, i seguenti 3 assiomi sono validi:

Ass.1:	$D_{xy} \geq 0$	non negatività
Ax.2:	$D_{xx} = 0$	riflessività
Ax.3:	$D_{xy} = D_{yx}$	simmetria

Una funzione $D: \quad X \times X \to \mathbb{R}$ is called a *quasi-distance* on X if, for all $x, y \in X$, solo i primi 2 assiomi valgono:

Ass.1: $D_{xy} \geq 0$ non negatività

Ax.2: $D_{xx} = 0$ riflessività

Una funzione $D: X \times X \to \mathbb{R}$ è chiamata una metrica su X se, per tutti $y, z \in X$, i seguenti 4 assiomi sono veri:

Ass.1: $D_{xy} \geq 0$ non negatività

Ax.2′: $D_{xy} = 0$ ⟺ $x=y$ forte riflessività

Ax.3: $D_{xy} = D_{yx}$ simmetria

Ax.4: $D_{xy} \leq D_{xz} + D_{zy}$ disuguaglianza triangolare

Le distanze che soddisfano solo i primi 3 assiomi sono chiamate distanze non metriche.

Una funzione $D: X \times X \to \mathbb{R}$ è chiamata una semi-metrica (o pseudo-metrica) su X se, per tutti $x, y, z \in X$, il i seguenti 4 assiomi sono validi:

Ass.1: $D_{xy} \geq 0$ non negatività

Ax.2: $D_{xx} = 0$ riflessività

Ax.3: $D_{xy} = D_{yx}$ simmetria

Ax.4: $D_{xy} \leq D_{xz} + D_{zy}$ disuguaglianza triangolare

Una funzione $D: X \times X \to \mathbb{R}$ si chiama un quasi-semi-metrico su X se, per tutti $x, y\ z \in X$, i seguenti 3 gli assiomi sono validi:

Ass.1: $D_{xy} \geq 0$ non negatività

Ax.2: $D_{xx} = 0$ riflessività

Ax.4: $D_{xy} \leq D_{xz} + D_{zy}$ disuguaglianza triangolare

Una funzione $D: X \times X \to \mathbb{R}$ si chiama una quasi-metrica su X se, per tutti $x, y\ z \in X$, i seguenti 3 assiomi valgono:

Ass.1: $D_{xy} \geq 0$ non negatività

Ax2′: $D_{xy} = 0$ ⟺ $x=y$ forte riflessività

$\text{Ax.4}: \quad D_{xy} \le D_{xz} + D_{zy}$ disuguaglianza triangolare

Una funzione $D: \quad X \times X \to \mathbb{R}$ è chiamato un ultra-metrico su X se, per tutti$x, y \; z \in X$, i seguenti 4 assiomi valgono:

Ass.1: $D_{xy} \ge 0$ non negatività

Ax2′: $D_{xy} = 0$ ..... *x=y* forte riflessività

$\text{Ax.3}: \quad D_{xy} = D_{yx}$ simmetria

$\text{Ax.4}': \quad D_{xy} \le \max\left\{D_{xz}, D_{zy}\right\}$ disuguaglianza ultra-metrica

L'assioma 2' è una condizione più forte dell'assioma 2, così come l'assioma 4' è più forte dell'assioma 4.

Le diverse classi di funzioni di dissimilarità possono essere distinte in base agli assiomi che soddisfano, come mostrato nella Figura 2 e nella Tabella 1. Si noti che tutte le funzioni di dissimilarità devono soddisfare almeno i due requisiti di base di non negatività (Ax. 1) e riflessività (Ax. 2) e sono ulteriormente distinte in distanze e quasi-distanze a seconda che la proprietà di simmetria (Ax. 3) sia soddisfatta o meno. Ovviamente, la classe delle quasi-distanze è la più ampia e include tutte le distanze.

Le distanze possono essere ulteriormente distinte in distanze metriche e distanze non metriche in base alla proprietà dell'ineguaglianza triangolare (Ax. 4): se l'ineguaglianza triangolare non è soddisfatta, allora una distanza è non metrica, altrimenti è metrica se è soddisfatta anche la proprietà della forte riflessività (Ax. 2'). Se la forte riflessività non è valida, cioè, possono esserci coppie di oggetti x ≠ y per cui Dxy = 0, allora una distanza non può essere considerata propriamente metrica ed è chiamata distanza semi-metrica. La classe delle quasi-metriche include tutte le quasi-distanze che soddisfano le proprietà dell'ineguaglianza triangolare (Ax. 4) e della forte riflessività (Ax. 2') e si differenziano dalle distanze metriche poiché le quasi-metriche non soddisfano la proprietà di simmetria (Ax. 3).Ovviamente, la classe delle quasi-metriche include tutte le distanze metriche.

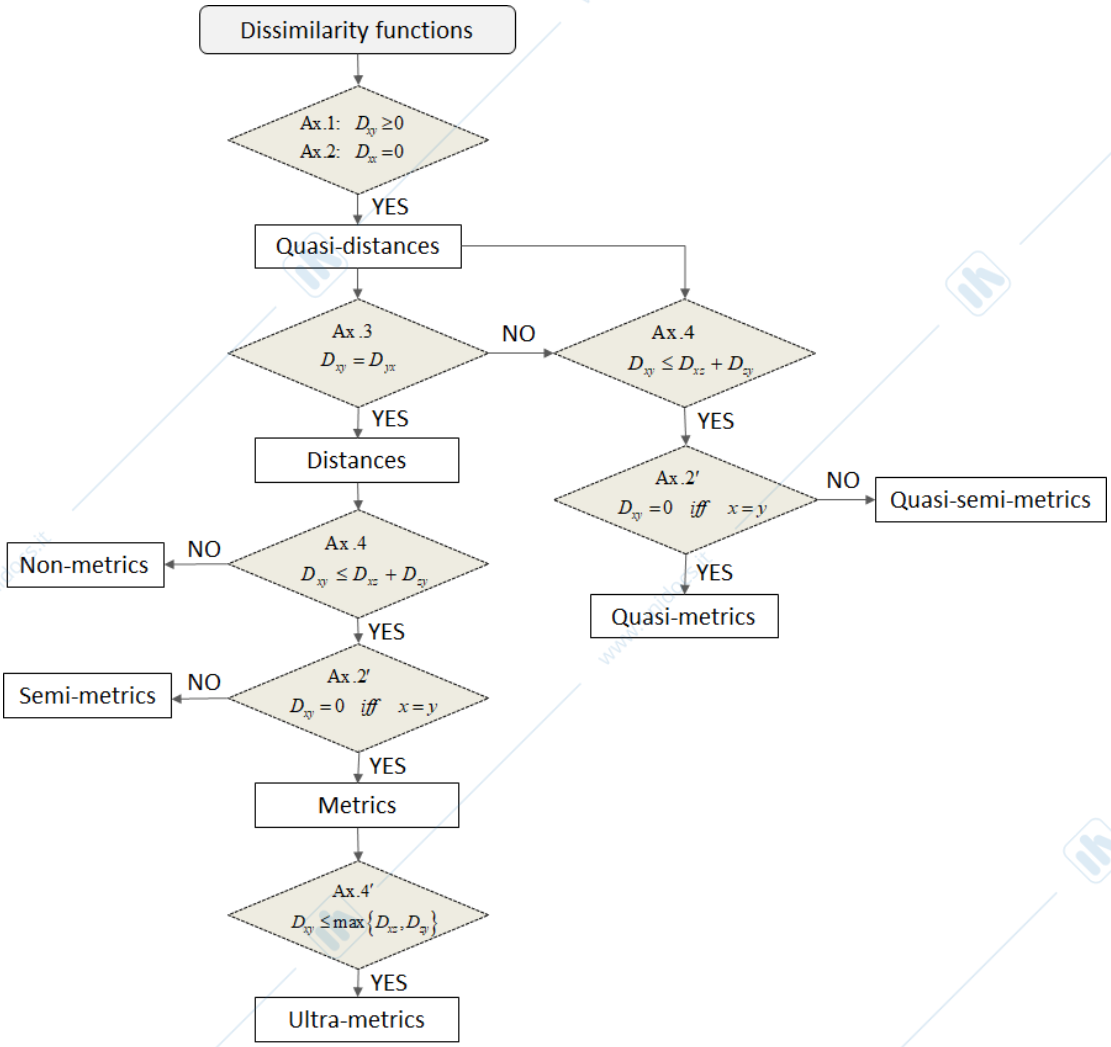

Figura 2 Diagramma di flusso delle relazioni tra le diverse classi di funzioni di dissimilarità.

Tabella 1. Diverse classi di misure di dissimilarità e i loro assiomi.

Funzione di dissimilarità	*Asse 1*	*Asse 2*	*Asse 2'*	*Asse 3*	*Ax. 4*	*Ax. 4'*
quasi-distanze	•	•				
distanze	•	•		•		
quasi-semi-metriche	•	•			•	
semi-metriche o pseudo-metriche	•	•		•	•	
quasi-metriche	•	•	•		•	
metriche	•	•	•	•	•	
ultra-metriche	•	•	•	•	•	•

2.3 Dalla distanza alla somiglianza

Una funzione S definita come $S : X \times X \to \mathbb{R}$ on X is called *similarity*, if, for all $x, y \in X$, i seguenti le proprietà sono:

Ass.1: $S_{xy} \geq 0$ non negatività

Ax.2: $S_{xx} = 0$ identità

Ax.3: $S_{xy} = S_{yx}$ simmetria

Le misure di similarità più comuni soddisfano anche una condizione di chiusura, cioè, $0 \leq S_{xy} \leq 1$. Il valore 1 indica la massima somiglianza, mentre il valore 0 indica la massima dissimilarità. In questo contesto, le misure di correlazione possono essere considerate un caso speciale delle misure di somiglianza; tuttavia, i valori di correlazione possono essere negativi, cioè compresi tra [-1, +1], e in questo caso non possono essere propriamente considerati come misure di somiglianza a causa della violazione del primo assioma che implica la non negatività.

La maggior parte delle misure di somiglianza può essere derivata dalle misure di distanza tramite funzioni di trasformazione appropriate, la cui scelta dipende principalmente dal fatto che la misura di distanza sia limitata o meno.

Alcune distanze, per definizione, sono intrinsecamente limitate al valore superiore di 1 e altre possono essere limitate nello stesso intervallo normalizzando il numero di variabili e/o applicando alcune procedure di scalatura.

Ad esempio, dato un dataset costituito da n oggetti e p variabili, dopo la scalatura dell'intervallo, qualsiasi distanza che varia tra 0 e p, può essere normalizzata tra 0 e 1 semplicemente facendo la media su p. Per queste distanze limitate [0, 1] ND, è possibile derivare le corrispondenti misure di similarità mediante le seguenti equazioni:

$$(1) \quad S_{xy}^{(1)} = 1 - D_{xy}^{N}$$

$$(2) \quad S_{xy}^{(2)} = 1 - \left(D_{xy}^{N}\right)^2$$

$$(3) \quad S_{xy}^{(3)} = \sqrt{1 - \left(D_{xy}^{N}\right)^2}$$

La misura di similarità ottenuta è naturalmente limitata a [0, 1], come di solito richiesto per una funzione di similarità.

Quando si trattano distanze illimitate, una procedura diretta per trasformarle in misure di similarità può essere ottenuta tramite funzioni di trasformazione specifiche concepite in modo tale che i valori di similarità risultino limitati tra 0 e 1. Le trasformazioni di similarità più conosciute per distanze illimitate D sono le seguenti:

$$(4) \quad S_{xy}^{(4)} = \frac{1}{1+D_{xy}}$$

$$(5) \quad S_{xy}^{(5)} = 1 - \frac{D_{xy}}{D_{\max}}$$

$$(6) \quad S_{xy}^{(6)} = \mathrm{e}^{-D_{xy}}$$

dove Dmax rappresenta la distanza massima tra tutte le possibili n x (n-1)/2 coppie di oggetti nel dataset.

La trasformazione di similarità di tipo (4) è la più semplice, ma il suo principale svantaggio è che la similarità è compressa verso zero quando sono presenti distanze significativamente maggiori rispetto a tutte le altre. Infatti, indipendentemente dalla procedura di scalatura adottata, la presenza di diverse variabili aumenta il valore della distanza per la maggior parte delle funzioni di distanza. Inoltre, questa trasformazione (4) non dovrebbe essere utilizzata per distanze normalizzate DN, la distanza massima in questo caso essendo pari a 1, il che a sua volta porterebbe alla similarità minima di 0.5 invece di 1. Per superare questo svantaggio, la trasformazione (4) dovrebbe essere ulteriormente implementata come:

$$S_{xy}^{(4')} = 2 \cdot \left(S_{xy}^{(4)} - 0.5 \right)$$

La funzione di trasformazione (6) soffre dello stesso svantaggio, il valore minimo essendo 0.368 (cioè, 1/e), che si ottiene quando la distanza normalizzata è pari al suo valore massimo di 1. In questo caso, dovrebbe essere adottata la seguente scalatura per avere valori che variano correttamente da 0 a 1:

$$S_{xy}^{(6')} = \frac{S_{xy}^{(6)} - 0.368}{1 - 0.368}$$

Per la funzione di trasformazione (5), è importante notare che ci sarà sempre almeno una coppia di oggetti x e y con similarità pari a zero, cioè la coppia di oggetti che ha la massima distanza sarà

hanno una similarità pari a 0 indipendentemente dal valore effettivo della loro distanza. Infatti, questa trasformazione di similarità fornisce valori di similarità relativi, a seconda della coppia di oggetti più distante.

Poiché la maggior parte delle misure di similarità sono derivate da misure di distanza tramite funzioni di trasformazione appropriate, anche le misure di dissimilarità D possono essere derivate dalle misure di similarità S utilizzando qualsiasi trasformazione monotonicamente decrescente di S.

Esempi di trasformazioni utilizzate per ottenere misure di dissimilarità dalle similarità sono:

$$D_{xy}^{(1)} = 1 - S_{xy}$$
$$D_{xy}^{(2)} = \sqrt{1 - S_{xy}}$$
$$D_{xy}^{(3)} = \sqrt{2 \cdot \left(1 - S_{xy}^2\right)}$$
$$D_{xy}^{(4)} = arccos\left(S_{xy}\right)$$
$$D_{xy}^{(5)} = -ln\, S_{xy}$$

dove Sxy si presume vari nell'intervallo [0, 1].

2.4 Pretrattamento dei dati

Quando si tratta di data mining, è spesso necessaria una pretrattazione dei dati per consentire un confronto equo delle variabili definite da diverse scale di misurazione e unità.

Il passo più rilevante della pretrattazione dei dati è rappresentato dalle procedure di scalatura dei dati, che consentono di comprimere tutte le variabili in una scala comparabile in modo che le misure di distanza/similarità tra gli oggetti possano sfruttare equamente le informazioni presenti in tutte le variabili, indipendentemente dalla loro scala numerica.

In generale, le misure di distanza definite per dati a valori reali richiedono che tutte le variabili siano comparabili, cioè definite nella stessa scala numerica. Se le variabili sono definite in scale diverse, le distanze tra gli oggetti subiscono effetti di distorsione semplicemente a causa delle diverse scale numeriche e non sono realmente influenzate da un contributo comparabile di tutte le variabili utilizzate per definire i dati. Pertanto, nel caso di dati a valori reali, la pretrattazione dei dati è necessaria prima del calcolo delle distanze.

Le procedure di scalatura dei dati eseguono la trasformazione di ciascuna variabile, separatamente, in una scala numerica comparabile. Sia n il numero totale di oggetti di un dataset e p il numero di variabili, xij sia il valore della j-esima variabile per l' i-esimo oggetto e x'_{ij} il valore scalato corrispondente. Le quantità Lj, Uj, jx¯ e js sono rispettivamente i valori minimo, massimo, medio e deviazione standard della j-esima variabile.

Pertanto, le procedure di scaling più comuni sono le seguenti:

- *Centralizzazione della media*

$$x'_{ij} = x_{ij} - \bar{x}_j \qquad \bar{x}'_j = 0$$

- *Scalare al massimo*

$$x'_{ij} = \frac{x_{ij}}{U_j} \qquad x'_{ij} \le 1 \quad U_j = \max_i \left(x_{ij} \right)$$

- *Scalatura dell'intervallo*

$$x'_{ij} = \frac{x_{ij} - L_j}{U_j - L_j} \qquad 0 \le x'_{ij} \le 1$$

$$L_j = \min_i \left(x_{ij} \right) \quad U_j = \max_i \left(x_{ij} \right)$$

- *Scalatura a varianza unitaria*

$$x'_{ij} = \frac{x_{ij}}{s_j^2} \qquad s'_j = 1$$

- *Autoscaling*

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad \bar{x}'_j = 0 \qquad s'_j = 1$$

- *scalatura di Pareto*

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j^2} \qquad \bar{x}'_j = 0 \qquad s'_j = \sqrt{s_j}$$

- *centratura logaritmica*

$$x'_{ij} = \log\left(x_{ij} \right) - \sum_{i=1}^{n} \log\left(x_{ij} \right) / n$$

La media e la deviazione standard dei dati scalati cambiano, per ogni j-esima variabile, secondo le seguenti relazioni:

$$\bar{x}'_j = \alpha \cdot \bar{x}_j + \beta \quad \text{e} \quad s'_j = \alpha \cdot s_j$$

dove α e β sono due parametri, che possono assumere valori diversi in base ai tipi di procedure di scalatura, come mostrato nella Tabella 2.

Tabella 2. Parametri α e β delle diverse procedure di scalatura.

Procedura di scalatura	α	β
Centralizzazione della media	1	$-\bar{x}_j$
Scalare al massimo	$1/U_j$	0
Scalatura dell'intervallo	$1/(U_j - L_j)$	$-L_j/(U_j - L_j)$
Scalatura a varianza unitaria	$1/s_j$	0
Autoscaling	$1/s_j$	$-\bar{x}_j / s_j$
scalatura di Pareto	$1/s_j^2$	$-\bar{x}_j / s_j^2$

Alcune procedure di scalabilità, come ad esempio il centraggio della media, non risolvono il problema della comparabilità delle variabili.

Il trattamento preliminare dei dati aggiuntivi è solitamente necessario per dati più complessi come spettri e dati composizionali, per i quali viene applicata preliminarmente la scalatura delle righe. Le procedure di scalatura delle colonne precedentemente definite vengono comunque eseguite dopo la scalatura preliminare delle righe. La Correzione Multiplicativa della Diffusione (MSC) e la Variabile Normale Standard (SNV) [14,15] sono due procedure di scalatura classiche per i dati spettrali, mentre la trasformazione del log-ratio è suggerita per i dati composizionali [16].

Esempio 1. Scalatura contro non scalatura

Questo semplice esempio mira a mostrare il ruolo e l'importanza della scalatura dei dati mentre si affronta l'analisi delle relazioni di somiglianza/diversità tra oggetti. Nella Tabella 3, i dati riguardanti 5 oggetti descritti da 3 variabili sono raccolti, nella scala originale e dopo il trattamento di scalatura dell'intervallo. Considera

due oggetti 1 e 2. Se la distanza euclidea classica fosse calcolata utilizzando le tre variabili originali, si otterrebbe un valore di distanza di 10,05. Se la distanza euclidea tra i due oggetti fosse calcolata considerando solo la prima variabile, il valore della distanza sarebbe 10 e questo non sarebbe molto diverso dal precedente. Ciò significa che la prima variabile da sola contribuisce fino al 99,5% del valore della distanza tra gli oggetti 1 e 2 e che le altre due variabili non influenzano molto il risultato. Se ora facciamo lo stesso calcolo considerando i valori scalati invece di quelli originali, la distanza euclidea tra gli oggetti 1 e 2 basata su tutte e 3 le variabili diventa 0.414 e il contributo al valore della distanza della prima variabile diminuisce al 12%. Pertanto, se le distanze vengono calcolate sul dataset grezzo in presenza di variabili con scale diverse, le variabili con grandi varianze (o scale numeriche) avrebbero il peso maggiore nel calcolo della distanza, nascondendo così il contributo delle variabili caratterizzate da varianze più piccole.

Tabella 3. Dati originali e scalati dell'Esempio 1 sul trattamento dei dati.

	Dati originali			*Dati scalati*		
Dati	x_1	x_2	x_3	x_1	x_2	x_3
1	100	2	0.2	0.714	0.333	0.400
2	90	1	0.1	0.571	0.000	0.200
3	120	4	0	1.000	1.000	0.000
4	70	3	0.5	0.286	0.667	1.000
5	50	3	0.3	0.000	0.667	0.600

3 Definizioni di misure di distanza e somiglianza

Le misure di distanza e somiglianza possono differire a seconda del tipo di dati a cui vengono applicate: dati a valori reali, dati binari, dati ordinati e dati di frequenza. Questi dati sono distinti in base alle variabili utilizzate per descrivere gli oggetti. Variabili come l'intensità del segnale, l'attività biologica, la pressione, la temperatura, la velocità di reazione, la concentrazione, i conteggi, sono misurate quantitativamente su una

scala intervallare o di rapporto; sono variabili quantitative e appartengono alla classe dei dati a valori reali.

Le variabili quantitative possono essere continue o discrete: nel primo caso, i valori che una variabile può assumere sono un insieme di valori infiniti o non numerabili; nel secondo caso, l'insieme dei valori è finito o numerabile e i valori sono solitamente interi. Variabili come colore, forma, texture, forniscono una classificazione degli oggetti in categorie che descrivono la qualità di un oggetto; queste variabili sono variabili qualitative e sono misurate su una scala nominale. Se queste variabili consentono un ordinamento o una classificazione degli oggetti, allora si dice che le variabili sono misurate su una scala ordinale. Se una variabile nominale consente solo due valori, ad esempio sì/no, presente/assente, acceso/spento, ecc., questa è definita come variabile binaria, che è solitamente codificata come 1/0 [1].

Nei paragrafi successivi, vengono introdotte misure di distanza/similarità sulla base dei tipi di dati per i quali sono definite.

3.1 Distanze per dati a valori reali

I dati a valori reali hanno tutte le variabili rappresentate da valori reali, come concentrazioni, intensità di segnale degli spettri, misure fisiche o chimiche quantitative. Diverse misure di dissimilarità sono state definite nella letteratura per dati a valori reali, le più conosciute sono le distanze euclidea, di Manhattan e di Mahalanobis.

Le distanze a valori reali possono essere suddivise in due classi principali in base all'intervallo che coprono: distanze illimitate, che sono distanze che vanno da zero all'infinito, e distanze limitate, che vanno da zero a un valore finito fisso, cioè distanze che hanno un valore massimo limitato da un limite superiore.

Le distanze illimitate più comuni sono raccolte nella Tabella 4, mentre le distanze limitate nella Tabella 5: Dxy è il simbolo generale per rappresentare la misura di dissimilarità tra qualsiasi coppia di oggetti x e y; i simboli xj e yj indicano i valori delle j-esime variabili per gli oggetti p-dimensionali x e y, rispettivamente. L'ultima colonna della Tabella 4 e della Tabella 5 include le formule per calcolare le misure di dissimilarità media i cui valori sono indipendenti dal numero p di variabili che descrivono gli oggetti.

Tabella 4. Distanze illimitate per dati a valori reali.

Eq.	*Distanza*	*Definizione*	*Intervallo*	*Media*
R1	*Euclidea*	$D_{xy}^{EUC} = \sqrt{\sum_{j=1}^{p}\left(x_j - s_j\right)^2}$	$0 \le D_{xy}^{EUC} < \infty$	$\bar{D}_{xy}^{EUC} = \frac{D_{xy}^{EUC}}{\sqrt{p}}$
R2	*Manhattan o isolato cittadino*	$D_{xy}^{MAN} = \sum_{j=1}^{p}\left\|x_j - s_j\right\|$	$0 \le D_{xy}^{MAN} < \infty$	$\bar{D}_{xy}^{MAN} = \frac{D_{xy}^{MAN}}{p}$
R3	*Lagrange*	$D_{xy}^{LAG} = \max_j \left\|x_j - s_j\right\|$	$0 \le D_{xy}^{LAG} < \infty$	$\bar{D}_{xy}^{LAG} = D_{xy}^{LAG}$
R4	*Minkowski*	$D_{xy}^{MIN} = \left[\sum_{j=1}^{p}\left\|x_j - s_j\right\|^q\right]^{1/q}$	$q > 0$ $0 \le D^{MIN} < \infty$	$\bar{D}^{MIN} = \frac{D^{MIN}}{p^{1/q}}$
R5	*Bhattacharyya*	$D_{xy}^{BHA} = \sqrt{\sum_{j=1}^{p}\left(\sqrt{x_j} - \sqrt{s_j}\right)^2}$	$x, y \ge 0$ $0 \le D_{xy}^{BHA} < \infty$	$\bar{D}_{xy}^{BHA} = \frac{D_{xy}^{BHA}}{\sqrt{p}}$
R6	*Mahalanobis*	$D_{xy}^{MAH} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}} \cdot \mathbf{S}^{-1} \cdot (\mathbf{x}-\mathbf{y})}$	$0 \le D_{xy}^{MAH} < \infty$	$\bar{D}_{xy}^{MAH} = \frac{D_{xy}^{MAH}}{\sqrt{p}}$
R7	*Mahalanobis centrato localmente*	$D_{xy}^{LCM} = \frac{1}{p} \cdot \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}} \cdot \mathbf{S}_{(s)}^{-1} \cdot (\mathbf{x}-\mathbf{y})}$	$0 \le D_{xy}^{LCM} < \infty$	$\bar{D}_{xy}^{LCM} = D_{xy}^{LCM}$
R8	*Forma di Penrose**	$D_{xy}^{PEN} = \sqrt{\sum_{j=1}^{p}\left[\left(x_j - \bar{x}\right) - \left(s_j - \bar{s}\right)\right]^2}$	$0 \le D_{xy}^{PEN} < \infty$	$\bar{D}_{xy}^{PEN} = \frac{D_{xy}^{PEN}}{\sqrt{p}}$

* $\bar{x}$ e $\bar{y}$ sono le due medie campionarie.

Una rappresentazione geometrica delle distanze illimitate più conosciute, cioè la distanza euclidea (R1), la distanza di Manhattan (R2) e la distanza di Lagrange (R3), è mostrata nella Figura 3. La distanza euclidea tra i punti x e y corrispondono al percorso più breve che unisce i due punti ($\sqrt{a^2+b^2}$); la Manhattan

La distanza, chiamata anche distanza taxi, è la somma dei percorsi più brevi lungo ciascuna dimensione (a + b); infine, la distanza di Lagrange, chiamata anche distanza di Chebyshev, è il percorso massimo tra i percorsi più brevi lungo ciascuna dimensione (a).

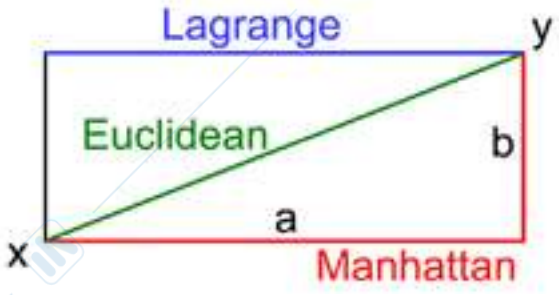

Figura 3 Rappresentazione geometrica delle distanze euclidea, di Manhattan e di Lagrange tra due punti x e y.

La distanza di Minkowski (R4) rappresenta una famiglia di misure di distanza, per cui maggiore è il valore di q, maggiore è l'importanza data alle grandi differenze. Le distanze euclidea (R1), di Manhattan (R2) e di Lagrange (R3) sono casi speciali della distanza di Minkowski (R4), corrispondenti a diversi valori dell'esponente q di questa distanza potenza: la distanza euclidea si ottiene per q = 2, Manhattan

distance for $q = 1$ and Lagrange distance for $q \to \infty$.

Le distanze di Minkowski per $0 < q < 1$ sono esempi di distanze non metriche [17]. La distanza non metrica è anche la distanza mediana, calcolata da un vettore p-dimensionale d delle distanze a coppie lungo ciascuna dimensione: $D_{xy}^{Med} = med_j\left(d_1, d_2, \ldots d_p\right)$, dove dj è la differenza tra i valori di x e y per la j-esima distanza; inoltre, la distanza k-mediana è calcolata scegliendo il k-esimo valore più basso del vettore delle distanze d.

Tra le misure di dissimilarità definite nella Tabella 4, la distanza di Mahalanobis (R6) è una funzione di dissimilarità relativa, che misura la dissimilarità tra due oggetti x e y non solo sulla base dei due oggetti considerati, ma tenendo anche conto delle informazioni sull'intera struttura del dataset, tramite la matrice di covarianza dei dati S. La distanza di Mahalanobis può essere considerata una misura di distanza affidabile quando esiste correlazione tra le variabili (cioè è in grado di sottovalutare le variabili correlate che forniscono altrimenti informazioni ridondanti).

Inoltre, si noti che la distanza di Mahalanobis è una versione estesa della distanza euclidea: se la matrice di covarianza S viene sostituita dalla matrice identità I, allora la distanza di Mahalanobis si riduce alla distanza euclidea.

La distanza di Mahalanobis DMAH è proporzionale al leverage, che è spesso utilizzato nel contesto della diagnostica di regressione, rilevamento di outlier e analisi del dominio di applicabilità, definito come:

$$h_{ic} = \mathbf{x}_i \cdot \left(\mathbf{X}_c^{\mathrm{T}} \mathbf{X}_c\right) \cdot \mathbf{x}_i^{\mathrm{T}} = \frac{\left(D_{ic}^{MAH}\right)^2}{n-1}$$

dove Xc è la matrice del modello centrata sul centroide c dei dati e n è il numero di oggetti nel dataset; hic misura la dissimilarità tra l'oggetto xi e il centroide dei dati, o, in altre parole, fornisce la "distanza" dell'i-esimo oggetto dal centro del modello rappresentato dalla matrice dei dati X.

Recentemente, è stata proposta in letteratura una variante della distanza di Mahalanobis, chiamata distanza di Mahalanobis localmente centrata (R7) [18]. Ciò che rende questa funzione diversa dalla classica distanza di Mahalanobis è il modo in cui viene valutata la dissimilarità tra gli oggetti x e y, poiché la matrice di covarianza dei dati è centrata in uno dei due oggetti invece che nel centroide dei dati. Ovviamente, si ottengono due valori diversi a seconda che la matrice di covarianza sia centrata in y (cioè, S(y)) o in x (cioè, S(x)), e pertanto, la proprietà di simmetria (Axiom 3) è violata con la conseguenza che la funzione di Mahalanobis localmente centrata non può essere considerata propriamente una distanza ma una quasi-distanza (cioè, LCM.

$$D_{xy} \neq D_{yx}^{LCM}).$$

Per rendere la Mahalanobis localmente centrata simmetrica e quindi conforme all'Axiom 3 delle distanze, possono essere applicate due procedure di simmetrizzazione:

$$D_{xy}^{MSA} = \frac{D_{xy}^{LCM} + D_{yx}^{LCM}}{2}$$

$$D_{xy}^{MSG} = \sqrt{D_{xy}^{LCM} \times D_{yx}^{LCM}}$$

il primo essendo la media aritmetica (MSA) dei due valori di dissimilarità e il secondo la loro media geometrica (MSG). Per semplicità, d'ora in poi, il simbolo MUxyD sarà utilizzato per indicare la funzione di Mahalanobis centrata localmente in x al posto del simbolo LCMxyD

, mentre

il simbolo D_{xy}^{ML} sostituirà il simbolo D_{yx}^{LCM} per indicare la funzione di Mahalanobis centrata localmente centrata in y. Si noti che il simbolo MU si riferisce agli elementi della sottmatrice triangolare superiore della matrice di dissimilarità contenente i valori LCM tra tutte le coppie di oggetti e il simbolo ML agli elementi della sottmatrice triangolare inferiore della stessa matrice. La distanza di forma di Pemrose (R8) è una distanza in cui ogni variabile è ridimensionata sulla media del campione. Nella distanza di Dehmer (R15), l'argomento del termine esponenziale corrisponde a variabili scalate a varianza unitaria [19].

Misure di distanza specifiche per dati a valori reali sono anche derivate da misure di similarità (Tabella 5). Tra queste, la distanza Jaccard-Tanimoto (R16) è derivata dal noto coefficiente di similarità Jaccard-Tanimoto SJT:

$$S_{xy}^{JT} = \frac{\sum_{j=1}^{p} x_j \cdot y_j}{\sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} s\grave{i}_j^2 - \sum_{j=1}^{p} x_j y_j}$$

e la distanza coseno (R19) è derivata dal coefficiente di similarità coseno SCC:

$$S_{xy}^{CC} = \frac{\sum_{j=1}^{p} x_j \cdot y_j}{\sqrt{\sum_{j=1}^{p} x_j^2 \cdot \sum_{j=1}^{p} s\grave{i}_j^2}}$$

Tabella 5. Distanze limitate su dati a valori reali. rxy è la correlazione di Pearson.

Eq.	*Distanza*	*Definizione*	*Intervallo*	*Media*
R9	*Canberra*	$D_{xy}^{CAN} = \sum_{j=1}^{p} \frac{\lvert x_j - s\grave{i}_j \rvert}{\lvert x_j \rvert + \lvert s\grave{i}_j \rvert}$	$0 \le D_{xy}^{CAN} \le p$	$\bar{D}_{xy}^{CAN} = \frac{D_{xy}^{CAN}}{p}$
R10	*Clark o coefficiente di divergenza*	$D_{xy}^{CLA} = \sqrt{\sum_{j=1}^{p} \left(\frac{x_j - s\grave{i}_j}{\lvert x_j \rvert + \lvert s\grave{i}_j \rvert} \right)^2}$	$0 \le D_{xy}^{CLA} \le p$	$\bar{D}_{xy}^{CLA} = \frac{D_{xy}^{CLA}}{\sqrt{p}}$
R11	*Wave-Edge*	$D_{xy}^{WE} = \sum_{j=1}^{p} \left(1 - \frac{\min(x_j, y_j)}{\max(x_j, y_j)} \right)$	$0 \le D_{xy}^{WE} \le p$	$\bar{D}_{xy}^{WE} = \frac{D_{xy}^{WE}}{p}$

R12	*Lance-Williams o Bray-Curtis*	$D_{xy}^{LW} = \frac{\sum_{j=1}^{p} \lvert x_j - s\grave{i}_j \rvert}{\sum_{j=1}^{p} \left(\lvert x_j \rvert + \lvert s\grave{i}_j \rvert \right)}$	$0 \le D_{xy}^{LW} \le 1$	$\bar{D}_{xy}^{LW} = D_{xy}^{LW}$
R13	*Soergel*	$D_{xy}^{SOE} = \frac{\sum_{j=1}^{p} \lvert x_j - s\grave{i}_j \rvert}{\sum_{j=1}^{p} \max\left(x_j, y_j\right)}$	$0 \le D_{xy}^{SOE} \le 1$	$\bar{D}_{xy}^{SOE} = D_{xy}^{SOE}$
R14	*Intersezione*	$D_{xy}^{INT} = 1 - \frac{\sum_{j=1}^{p} min\{x_i, y_i\}}{min\left\{\sum_{j=1}^{p} x_j, \sum_{j=1}^{p} s\grave{i}_j\right\}}$	$0 \le D_{xy}^{INT} \le 1$	$\bar{D}_{xy}^{INT} = D_{xy}^{INT}$
R15	*Dehmer*	$D_{xy}^{DEM} = p - \sum_{j=1}^{p} e^{-\left(\frac{x_j - s\grave{i}_j}{s_j}\right)^2}$	$0 \le D_{xy}^{DEM} \le p$	$\bar{D}_{xy}^{DEM} = \frac{D_{xy}^{DEM}}{p}$
R16	*Jaccard-Tanimoto*	$D_{xy}^{JT} = \sqrt{1 - \frac{\sum_{j=1}^{p} x_j \cdot y_j}{\sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} s\hat{i}_j^2 - \sum_{j=1}^{p} x_j \cdot y_j}}$	$0 \le D_{xy}^{JT} \le 1$	$\bar{D}_{xy}^{JT} = D_{xy}^{JT}$
R17	*Pearson*	$D_{xy}^{PEA} = 1 - r_{xy}$	$0 \le D_{xy}^{PEA} \le 2$	$\bar{D}_{xy}^{PEA} = D_{xy}^{PEA}$
R18	*Correlazione*	$D_{xy}^{COR} = \frac{1 - r_{xy}}{2}$	$0 \le D_{xy}^{COR} \le 1$	$\bar{D}_{xy}^{COR} = D_{xy}^{COR}$
R19	*Coseno o angolare*	$D_{xy}^{CD} = 1 - \frac{\sum_{j=1}^{p} x_j \cdot y_j}{\sqrt{\sum_{j=1}^{p} x_j^2 \cdot \sum_{j=1}^{p} s\hat{i}_j^2}}$	$0 \le D_{xy}^{CD} \le 1$	$\bar{D}_{xy}^{CD} = D_{xy}^{CD}$
R20	*sferico* *	$D_{xy}^{SPH} = arc\cos\left(\mathbf{x}' \cdot \mathbf{y}'\right)$	$0 \le D_{xy}^{SPH} \le \pi$	$\bar{D}_{xy}^{SPH} = \frac{D_{xy}^{SPH}}{\pi}$

* $x_j' = \frac{x_j}{\sqrt{\sum_{j=1}^{p} x_j^2}}$ $\quad s\hat{i}_j = \frac{s\grave{i}_j}{\sqrt{\sum_{j=1}^{p} s\hat{i}_j^2}}$

È interessante notare che le distanze euclidee e Jaccard-Tanimoto sono intrinsecamente correlate.

Infatti, la distanza euclidea al quadrato può essere riscritta come:

$$\left(D_{xy}^{EUC}\right)^2 = \sum_{j=1}^{p}\left(x_j - s\grave{i}_j\right)^2 = \sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} s\hat{i}_j - 2\cdot\sum_{j=1}^{p} x_j y_j$$

e la distanza Jaccard-Tanimoto al quadrato come:

$$\left(D_{xy}^{JT}\right)^2 = 1 - \frac{\sum_{j=1}^{p} x_j y_j}{\sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} s\hat{i}_j - \sum_{j=1}^{p} x_j y_j} = \frac{\sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} s\hat{i}_j - 2\cdot\sum_{j=1}^{p} x_j y_j}{\sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} s\hat{i}_j - \sum_{j=1}^{p} x_j y_j} = \frac{\left(D_{xy}^{EUC}\right)^2}{\sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} s\hat{i}_j - \sum_{j=1}^{p} x_j y_j}$$

Secondo questa relazione, deriva che la distanza Jaccard-Tanimoto al quadrato può essere vista come una distanza euclidea al quadrato normalizzata.

Le distanze di Pearson (R17) e di correlazione (R18), solitamente applicate per misurare la correlazione tra variabili ma qui applicate a coppie di oggetti, sono derivate dal coefficiente di correlazione di Pearson rxy, che è la misura di correlazione bivariata più conosciuta e stima il grado di associazione tra due oggetti x e y, definita come:

$$r_{xy} = \frac{\sum_{j=1}^{p}\left(x_j - \bar{x}\right)\cdot\left(s\grave{i}_j - \bar{s}\grave{i}\right)}{\sqrt{\sum_{j=1}^{p}\left(xx_j - \bar{}\right)^2 \cdot \sum_{j=1}^{p}\left(s\grave{i}_j - \bar{s}\grave{i}\right)^2}} \qquad -1 \le r_{xy} \le +1$$

dove x ̄e y sōno le medie dei vettori x e y, rispettivamente.

Dal coefficiente di correlazione, è stata anche definita la distanza di Pearson quadrata come:

$$D_{xy}^{SQP} = 1 - r_{xy}^2$$

dove coppie di oggetti con correlazione uguale a entrambi – 1 o +1 sono considerate simili tra loro ($D_{xy}^{SQP} = 0$).

Per ottenere una misura adeguata e limitata nell'intervallo [0, 1], la distanza di Pearson (R17) è stata scalata, dando la distanza di correlazione (R18).

$$D_{xy}^{COR} = \frac{D_{xy}^{PEA}}{2}$$

Va notato che la distanza di correlazione, a differenza delle distanze classiche, non tiene conto delle differenze sistematiche tra gli oggetti, poiché misura l'associazione tra i profili degli oggetti.

Inoltre, quando gli oggetti sono descritti da sole due variabili, la distanza di correlazione dà sempre un valore pari a 1 o -1, proiettando così tutte le distanze solo in due punti, e non può essere calcolata per dati descritti da una sola variabile.

Infine, l'uso nella chemometria della distanza sferica (R20) è stato proposto in [20].

Una misura di distanza è invariabile rispetto alla scala se la seguente relazione è soddisfatta:

$$D(x,y) = D(\alpha x + \beta, \alpha y + \beta)$$

dove α e β sono i parametri di scala. Più specificamente, una distanza ha la proprietà di:

1) invarianza alla traslazione, cioè invarianza rispetto alla traslazione rispetto all'origine, se:

$$D(x,y) = D(x+\beta, y+\beta) \qquad \beta \geq 0$$

con α = 1 nell'espressione di invarianza di scala;

2) invarianza di scala, cioè invarianza rispetto alla dilatazione, se:

$$D(x,y) = D(\alpha x, \alpha y) \qquad \alpha \geq 1$$

con β = 0 nell'espressione di invarianza di scala.

Tra le distanze raccolte nelle Tabelle 4 e 5, le distanze classiche di Mahalanobis, Mahalanobis localmente centrato, Dehmer, Pearson e le distanze di correlazione sono scalari invarianti poiché entrambe le condizioni (cioè, invariabilità di traduzione e invariabilità di scala) sono soddisfatte. Euclidea, Manhattan, Lagrange, Minkowski e Penrose sono invarianti rispetto alla traduzione, poiché qualsiasi costante aggiunta sia a x che a y scompare quando viene calcolata la differenza tra i valori, mentre non sono invarianti rispetto alla dilatazione. Al contrario, la maggior parte delle distanze limitate della Tabella 5 risulta essere invarianti rispetto alla dilatazione ma non alla traduzione, poiché si basano sul rapporto di due quantità.

Esempio 2. Proprietà di invarianza delle distanze.

Le proprietà di invarianza delle distanze sono state ulteriormente investigate attraverso un semplice esercizio.

Considera due oggetti x e y descritti da variabili scalate nell'intervallo [0, 1]. Quindi, il parametro di spostamento

β viene aggiunto a tutti i valori variabili di x e y e varia tra 0 e 5 con un passo di 0,5. Per ogni diverso valore del parametro di traslazione, la distanza tra i due oggetti x e y viene calcolata utilizzando le diverse funzioni di distanza delle Tabelle 4 e 5. Questo esercizio può essere ripetuto, mantenendo il parametro di traslazione β costante a valore 0 e moltiplicando per il parametro di dilatazione α le variabili di entrambi gli oggetti x e y. Il parametro di dilatazione α varia da 1 a 2 con un passo di 0,2 e, per ogni diverso valore, vengono calcolate le distanze delle Tabelle 4 e 5. Infine, i risultati di questo calcolo sono mostrati nelle Figure 4 e 5, dove è facile vedere quanto le diverse funzioni di distanza siano sensibili ai parametri di traslazione e dilatazione.Le distanze di Canberra, Lance-Williams, Clark, Soergel, Wave-Edge, Jaccard-Tanimoto, Coseno, Intersezione, Sferica e Bhattacharyya non sono chiaramente invarianti rispetto alla traslazione (Figura 4). Per queste funzioni di distanza, infatti, l'origine degli assi è significativa e quindi, se i due oggetti vengono allontanati dall'origine degli assi, la distanza tra di essi tende verso zero.

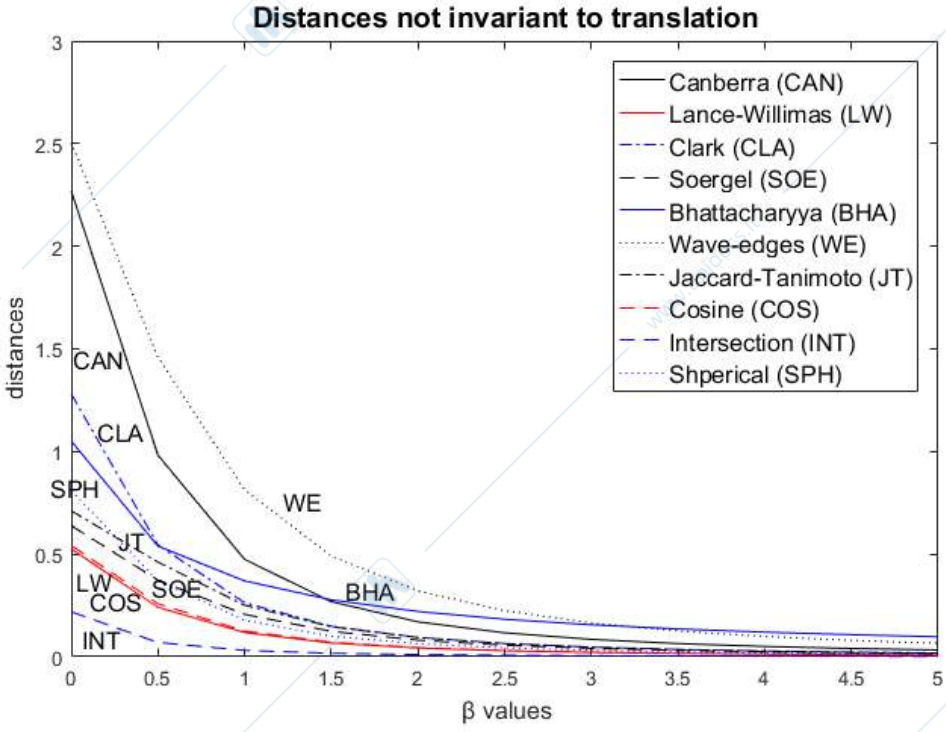

Figura 4 Profili dei diversi valori di distanza tra due oggetti ottenuti per valori crescenti del parametro di traslazione.

Le distanze euclidea, manhattaniana, di Lagrange, di Penrose e di Bhattacharyya non soddisfano la proprietà di invarianza rispetto alla dilatazione e, come si può vedere nella Figura 5, tendono ad aumentare significativamente man mano che il valore di α aumenta: le distanze euclidea, manhattaniana, di Penrose e di Lagrange aumentano di un fattore α, le distanze di Bhattacharyya di un fattore $\sqrt{\alpha}$.

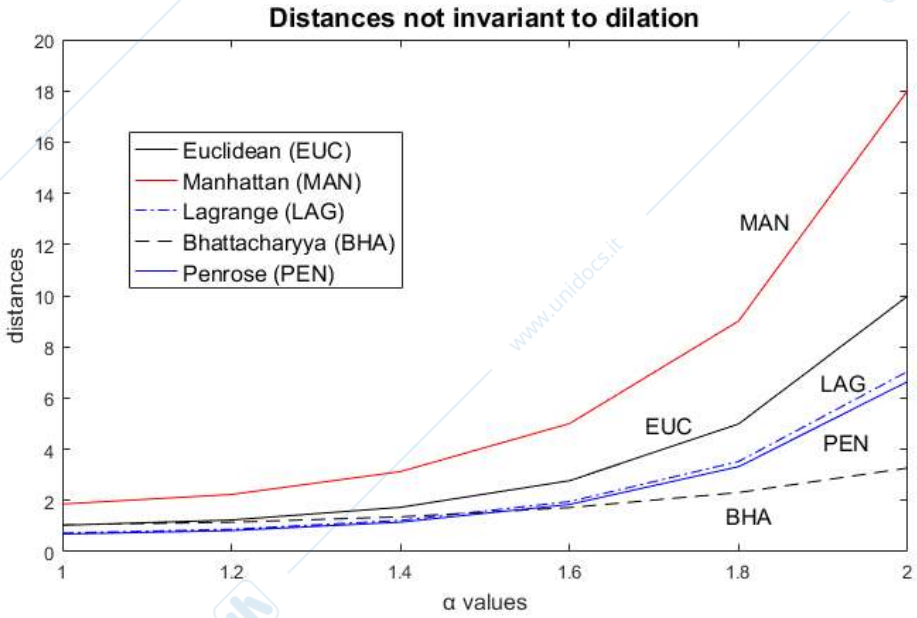

Figura 5 Profili dei diversi valori di distanza tra due oggetti ottenuti per valori crescenti del parametro di dilatazione.

Tra le distanze considerate, l'unica distanza che non soddisfa entrambe le proprietà di invarianza è la distanza di Bhattacharyya, sebbene con una bassa sensibilità sia ai parametri di traslazione che di dilatazione.

3.2 Distanze ponderate

Le distanze ponderate si ottengono pesando ciascuna variabile j-esima con un peso definito dall'utente wj solitamente sotto il vincolo:

$$\sum_{j=1}^{p} w_j = 1$$

Si può notare che se si desidera avere tutte le variabili con la stessa importanza sulla distanza, allora tutti i pesi avrebbero lo stesso valore $w_j = 1/p$.

In altri casi, i pesi possono essere definiti dall'utente e stabiliti da ulteriori informazioni a priori sulla variabile.

Esempi semplici di distanze pesate sono:

$$D_{xy}^{wEUC} = \sqrt{\sum_{j=1}^{p} w_j \cdot \left(x_j - s_j\right)^2}$$

$$D_{xy}^{wMAN} = \sum_{j=1}^{p} w_j \cdot \left| x_j - s_j \right|$$

Si può notare che, ad esempio, la distanza proposta da Decheng et al. [21]

$$D_{xy} = \sqrt{\sum_{j=1}^{p} w_j \cdot \left(\frac{x_j - s_j}{s_j}\right)^2}$$

è una distanza euclidea pesata calcolata su dati scalati a varianza unitaria.

3.3 Distanze per dati ordinati

Quando si tratta di dati ordinali, è implicito che i dati possano essere ordinati e a ciascuna entità possa essere assegnato un rango; per questo tipo di dati sono necessarie misure specifiche di distanza/similarità.

Un ranking (o permutazione) è qualsiasi vettore $r\{r_1, r_2, \ldots, r_n\}$ con tutti i ri che sono numeri diversi da $\{1,2,\ldots,n\}$. Vettori di ranking speciali sono n corrispondenti a un ordine naturale (o ordine di identità) 1, 2,..., n (e.g. alphabetic, age, etc.) and π corrispondente a qualche permutazione di riferimento di n.

The most common association measures between two rankings or permutations are the *Spearman* ρ *coefficient* (or *Spearman rank coefficient* ρ) and the *Kendall* τ *coefficient* [22]. The *Spearman* ρ il coefficiente per due vettori classificati r e s è definito come:

$$\rho_{rs} = 1 - \frac{6 \cdot \sum_{i=1}^{n} \left[r_i - s_i\right]^2}{n^3 - n} \qquad -1 \le \rho_{rs} \le +1$$

dove ri e si indicano i ranghi dell'oggetto i-esimo nell'intervallo [1, n]. Questa espressione della correlazione di rango di Spearman corrisponde alla correlazione di Pearson calcolata sui ranghi, cioè

$$\rho_{rs} = \frac{\sum_{i=1}^{n}[r_i - \bar{r}]\cdot[s_i - \bar{s}]}{\sqrt{\sum_{i=1}^{n}[r_i - \bar{r}]^2 \cdot \sum_{i=1}^{n}[s_i - \bar{s}]^2}}$$

dove $\bar{r}$ e $\bar{s}$ sono i ranghi medi dei due vettori r e s.

La distanza di Spearman è la distanza euclidea sui ranghi, che è definita come:

$$D_{rs}^{SPE} = \sqrt{\sum_{i=1}^{n}(r_i - s_i)^2}$$

La distanza di Spearman footrule è la distanza di ranking di Manhattan ed è definita come:

$$D_{rn}^{SFR} = \sum_{i=1}^{n}|r_i - s_i|$$

Nella letteratura più recente, la distanza di Spearman footrule è stata chiamata Somma delle Differenze di Ranking (SRD) ed è stata applicata per confronti di modelli e metodi; è definita come [23,24,25]:

$$D_{r\pi}^{SRD} = \sum_{i=1}^{n}|r_i - \pi_i|$$

where $\boldsymbol{r}$ are the actual ranks of the n objects and $\boldsymbol{\pi}$ è una permutazione di riferimento o la permutazione nulla $\boldsymbol{n}$.

The *Kendall* τ il coefficiente è una misura di correlazione tra i ranghi definita come:

$$\tau = \frac{2}{n\cdot(n-1)}\cdot\sum_{1\le i<j\le n} sgn(r_i - r_j)\cdot sgn(s_i - s_j) \qquad -1 \le \tau \le +1$$

dove sgn è la funzione segno.

La distanza di Kendall D_{rn}^{KEN} è definito come:

$$D_{rs}^{KEN} = \frac{2}{n\cdot(n-1)}\cdot\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\delta_{ij}$$

$$\delta_{ij} = \begin{cases} 1 & se\ (r_i > r_j) \wedge (s_i < s_j) \ \vee \ (r_i < r_j) \wedge (s_i > s_j) \\ 0 & altrimenti \end{cases}$$

corrispondente al numero totale di inversioni a coppie.

Tra la regola di Spearman e la distanza di Kendall, si mantiene la seguente relazione:

$$D_{rn}^{KEN} \le D_{rn}^{SFR} \le 2 \cdot D_{rn}^{KEN}$$

Altre possibili misure di distanza sui ranghi sono la distanza di Mahalanobis, definita come:

$$D_{rs}^{DMR} = 2 \cdot \sum_{i=1}^{n} \frac{\left(r_i - s_i\right)^2}{\left(r_i + s_i\right)}$$

e la distanza di Bhattacharyya, definita come:

$$D_{rs}^{BAH} = \sqrt{\sum_{i=1}^{n} \left(\sqrt{r_i} - \sqrt{s_i}\right)^2}$$

Esempio 3. Misure di somiglianza/diversità per dati classificati.

Per spiegare meglio come calcolare le misure di somiglianza/diversità per dati classificati, qui viene riportato un semplice esempio. Supponiamo di avere due classifiche a 4 dimensioni r = {3, 4, 2, 1} e s = {3, 1, 4, 2}. Quindi, le misure di somiglianza/diversità sopra menzionate vengono calcolate come:

Spearman ρcoefficiente:

$$\rho_{rs} = 1 - \frac{6 \cdot \left(0^2 + 3^2 + 2^2 + 1^2\right)}{4^3 - 4} = -0.4$$

Spearman ρdistanza:

$$D_{rs}^{SPE} = \left(0^2 + 3^2 + 2^2 + 1^2\right) = 14$$

Distanza di rango di Mahalanobis:

$$D_{rs}^{MRD} = 2 \cdot \left(0 + 1.80 + 0.67 + 0.33\right) = 5.6$$

Distanza di Bhattacharyya:

$$D_{rs}^{BHA} = \sqrt{\left(\sqrt{3} - \sqrt{3}\right)^2 + \left(\sqrt{4} - \sqrt{1}\right)^2 + \left(\sqrt{2} - \sqrt{4}\right)^2 + \left(\sqrt{1} - \sqrt{2}\right)^2} = 1.23$$

Kendall τcoefficiente:

$$\tau_{rs} = \frac{2 \cdot \left(-1 - 1 + 1 - 1 - 1 + 1\right)}{4 \cdot \left(4 - 1\right)} = \frac{2 \cdot \left(-2\right)}{12} = -0.33$$

Kendall τdistanza:

$$D_{rs}^{KEN} = \frac{2}{4\cdot(4-1)}\cdot\left(1_{(1,2)}+1_{(1,3)}+1_{(2,3)}+1_{(2,4)}\right)=0.67$$

3.4 Distanze per dati di frequenza

Una distribuzione di frequenza è la distribuzione del numero totale di osservazioni tra le varie categorie di una variabile. Data una distribuzione di frequenza, il termine f(m) rappresenta la frequenza osservata nella categoria m della variabile in studio per l'analisi di bontà di adattamento o la frequenza congiunta di due categorie delle variabili.

La misura più generale tra le distribuzioni di frequenza è la metrica Lp definita come:

$$D^{Lp} = \left[\sum_m \left|f_1(m)-f_2(m)\right|^k\right]^{1/k} \qquad k \geq 1$$

dove f1(m) e f2(m) sono le frequenze osservate di due distribuzioni di frequenza; questa misura è l'analogo della distanza di Minkowski (R4) sui dati a valore reale (Tabella 4).

La statistica chi-quadrato (o statistica χ2) è utilizzata come misura del grado di associazione tra le distribuzioni delle variabili categoriche[26] utilizzando valori interi delle frequenze. È definita come:

$$\chi^2 = \sum_m \frac{\left(f(m)-\hat{f}(m)\right)^2}{\hat{f}(m)}$$

dove f(m) e $\hat{f}(m)$ sono le frequenze osservate e attese, rispettivamente. Il chi-quadrato statistic is usually calculated from a contingency table of r rows and c columns $(r\times c)$, che coinvolge un elenco di tutti i valori osservati delle variabili in studio e quante volte ciascun valore è osservato.

La frequenza attesa nella tabella di contingenza è calcolata come:

$$\hat{f}_{ij} = \frac{R_i \cdot C_j}{n}$$

dove Ri è la somma della i-esima riga delle frequenze osservate e Cj è la somma della j-esima colonna delle osservate frequencies; n is the total number of observations and the degrees of freedom are $(r-1)\cdot(c-1)$.

La versione simmetrica della distanza χ2 è una quasi-distanza definita come [5]:

$$D^{\chi} = 2 \cdot \sum_{m} \frac{\left(f(m) - \hat{f}(m)\right)^2}{f(m) + \hat{f}(m)}$$

Esempio 4. χ2 misure per dati di frequenza.

Un esempio di χ^2 buona adattabilità per dati costituiti da 100 sostanze con un'attività teorica il rapporto di 3:1 è riportato di seguito.

	attivo	non attivo	
f osservato	84	16	$n = 100$
^f atteso	*(75)*	*(25)*	

$$\chi^2 = \frac{(84-75)^2}{75} + \frac{(16-25)^2}{25} = 4.32 \quad \text{e} \quad D^{\chi} = 2 \cdot \left[\frac{(84-75)^2}{84+75} + \frac{(16-25)^2}{16+25} \right] = 4.98$$

Un esempio per testare l'indipendenza di due variabili categoriche (colore dei capelli e sesso in humans) by a χ^2 la statistica è riportato di seguito.

Sesso \ Colore dei capelli	Nero	Marrone	Biondo	Rosso	Totale
Maschio	32	43	16	9	$R_1 = 100$
	(29.0)	*(36.0)*	*(26.7)*	*(8.3)*	
Femmina	55	65	64	16	$R_2 = 200$
	(58.0)	*(72.0)*	*(33.3)*	*(16.7)*	
	$C_1 = 87$	$C_2 = 108$	$C_3 = 80$	$C_4 = 25$	$n = 300$

La prima frequenza attesa è calcolata come:

$$\hat{f}_{11} = \frac{R_1 \cdot C_1}{n} = \frac{100 \times 87}{300} = 29.0$$

Poi, $\chi^2 = 9.03$ e $D^{\chi} = 2 \cdot \left[\frac{(32-29)^2}{32+29} + \frac{(43-36)^2}{43+36} + \ldots + \frac{(16-16.7)^2}{16+16.7} \right] = 9.81$.

Ancora importante per misurare le differenze tra le distribuzioni di frequenza è la divergenza di Kulback-Leibler (o rapporto di verosimiglianza logaritmica), che è definita come:

$$D^{KL} = \sum_m f(m) \cdot ln \frac{f(m)}{\hat{f}(m)}$$

Poiché la divergenza DKL non è una distanza adeguata a causa della sua asimmetria, una versione simmetrica è definita come:

$$D^{KLS} = \frac{1}{2} \cdot \left[\sum_m f(m) \cdot ln \frac{f(m)}{\hat{f}(m)} + \sum_m \hat{f}(m) \cdot ln \frac{\hat{f}(m)}{f(m)} \right]$$

Inoltre, due misure di similarità note per i dati di frequenza sono derivate dalla similarità di Tanimoto e dalla similarità di Tanimoto modificata, rispettivamente definite come:

$$S^T = \frac{\sum_{j=1}^{p} \min\left(f_{1j}, f_{2j}\right)}{\sum_{j=1}^{p} f_{1j} + \sum_{j=1}^{p} f_{2j} - \sum_{j=1}^{p} \min\left(f_{1j}, f_{2j}\right)}$$

e

$$S^{MT} = \frac{2 \cdot \sum_{j=1}^{p} \min\left(f_{1j}, f_{2j}\right)}{\sum_{j=1}^{p} f_{1j} + \sum_{j=1}^{p} f_{2j}}$$

dove f1j è il numero di occorrenze del j-esimo evento (la variabile) per il primo caso e f2j è il numero di occorrenze dello stesso j-esimo evento per il secondo caso. Le misure di distanza corrispondenti sono derivate come

$$D = 1 - S .$$

Infine, la probabilità di errore nella classificazione (ER), per una classificazione a due classi nel contesto di the Bayes rule, with a-priori probability π and 1 - π, e le corrispondenti probabilità di densità p1 e p2, possono essere trasformate in una distanza. In particolare, la distanza bayesiana tra due distribuzioni di classe può essere definita come:

$$D^{BAIA} \equiv 1 - ER = 1 - \sum_m min\left[\pi \cdot p_1(m), (1-\pi) \cdot p_2(m)\right]$$

3.5 Misure di similarità binaria

Le variabili binarie i cui valori sono uno o zero (presenza o assenza, sì/no, attivo/non attivo, ecc.) sono largamente comuni nell'analisi dei dati perché queste variabili sono in grado, ad esempio, di descrivere la presenza/assenza di un segnale a una determinata lunghezza d'onda di uno spettro, la presenza/assenza di un gruppo funzionale specifico o di un frammento molecolare in una molecola, di definire se un composto è attivo o inattivo rispetto a un certo obiettivo farmacologico e, in generale, se una certa caratteristica o attributo è osservato o meno.

Per trattare le variabili binarie, diversi coefficienti di similarità sono stati proposti in letteratura dalla fine del XVIII secolo e possono tutti essere descritti come segue. Siano due oggetti descritti dai vettori binari x e y, ciascuno composto da p variabili con valori 0/1. I coefficienti di associazione comuni sono calcolati dai dati riportati in una tabella di frequenza (Figura 6), dove a, b, c e d sono le frequenze degli eventi (x = 1 e y = 1), (x = 1 e y = 0), (x = 0 e y = 1) e (x = 0 e y = 0), rispettivamente, nella coppia di vettori binari che descrivono i due oggetti; p è il numero totale di attributi (cioè, variabili), uguale a a + b + c + d, che è la lunghezza di ciascun vettore binario.

	$y = 1$	$y = 0$	
$x = 1$	a	b	$a + b$
$x = 0$	c	d	$c + d$
	$a + c$	$b + d$	p

Figura 6 Tabella di frequenza delle quattro possibili combinazioni di valori 0 e 1 per due campioni binari x e y.

La tabella di frequenza può essere letta come segue: a è il numero di "presenze comuni" degli attributi e d il numero di "assenze comuni" in x e y; a + b è il numero di attributi presenti in x e a + c il numero di attributi presenti in y. Le voci diagonali a e d forniscono quindi informazioni su

la somiglianza tra i due vettori, mentre le voci b e c forniscono informazioni sulla loro dissimilarità.

Esempio 5. Tabella di frequenza per dati binari.

Un semplice esempio è presentato di seguito per mostrare come vengono calcolate le frequenze a, b, c e d. Siano due oggetti rappresentati dai vettori x e y, ciascuno descritto da dieci variabili binarie (cioè, p = 10):

$$\mathbf{x}: \ 1 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1$$
$$\mathbf{y}: \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1$$

Quindi, a, b, c e d assumono i seguenti valori (Figura 7):

	$y = 1$	$y = 0$	
$x = 1$	4	2	6
$x = 0$	1	3	4
	5	5	10

Figura 7 Valori dei coefficienti a, b, c e d della tabella di contingenza.

I coefficienti di similarità binaria sono solitamente distinti in 1) coefficienti simmetrici che utilizzano sia a che d, cioè, lo stato di doppio zero (d) per due oggetti è trattato esattamente allo stesso modo di qualsiasi altra coppia di valori e dovrebbe essere utilizzato quando lo stato zero è una base valida per confrontare due oggetti; 2) coefficienti asimmetrici, al contrario, ignorano tali attributi a doppio zero nel calcolo della similarità;

3) coefficienti basati sulla correlazione, che sono definiti nell'intervallo [-1, +1], tengono conto della differenza tra la frequenza di occorrenza delle concordanze (cioè, ad) e la frequenza di occorrenza delle discordanze (cioè, bc).

Tutti i coefficienti di similarità binaria proposti in letteratura sono elencati nella Tabella 6 [13]. La maggior parte di questi coefficienti è naturalmente definita nell'intervallo [0, 1]. Per quei coefficienti che hanno intervalli diversi da [0, 1], possono essere ridimensionati utilizzando la seguente trasformazione lineare:

$$S' = \frac{S + \alpha}{\beta}$$

dove S è il valore di similarità originale, S' è la funzione ridimensionata nell'intervallo [0, 1], e α e β sono parametri numerici i cui valori sono riportati nella Tabella 6 (dove, ovviamente, α = 0 e β = 1 indica che non è necessaria alcuna trasformazione per ottenere l'intervallo desiderato). La Tabella 6 raccoglie anche le condizioni matematiche che devono essere applicate per rendere ciascun coefficiente binario valido per qualsiasi combinazione delle frequenze a, b, c e d.

I coefficienti di similarità binaria più comuni sono il coefficiente di Jaccard-Tanimoto (B3) che enfatizza la presenza di caratteristiche comuni a, trascurando l'assenza di caratteristiche comuni d e il semplice matching (B1) che tiene conto sia della presenza che dell'assenza di caratteristiche comuni.

Una versione pesata del coefficiente di Jaccard-Tanimoto (B3) è il coefficiente di similarità di Tversky, definito come:

$$S_{xy}^{TV} = \frac{a}{a + \gamma \cdot b + \delta \cdot c} \qquad 0 \le S_{xy}^{TV} \le 1$$

where γ and δ are user-defined parameters. In particular, equal values of γ e δ forniscono un contributo simmetrico alle due frequenze di dissimilarità b e c, come, ad esempio, nel Jaccard-Tanimoto coefficient (B3), for which γ = δ = 1, in the Gleason coefficient (B4) when γ = δ = 1/2; nel Sokal-Sneath coefficient (B12) when γ = δ = 2 and in Jaccard coefficient (B14) when γ = δ = 1/3; diverso values of γ and δ forniscono un contributo asimmetrico, come, ad esempio, nel Dice-Wallace, Post-*Snijders coefficient* (B31), for which δ = 1 and γ = 0; questo coefficiente può essere interpretato come la frazione dell'oggetto x che è in comune con l'oggetto y.

Il coefficiente CT5 (B43) è la versione normalizzata di una misura derivata da un'analisi bayesiana della probabilità, che è stata definita come [27]:

$$w_{xy} = \ln \frac{\mathrm{p}(x=1 \wedge si=1) \times \mathrm{p}(x=0 \wedge si=0)}{\mathrm{p}(x=0 \wedge si=1) \times \mathrm{p}(x=1 \wedge x=0)} = \ln\left[\frac{ad}{bc}\right] \qquad -\infty < w_{xy} < +\infty$$

dove p è la probabilità dell'evento.

Tabella 6. Coefficienti di similarità binaria. Nella colonna "Condizioni", den indica il denominatore della funzione.

Eq.	*Coefficiente di similarità*	*Definizione*	α β	*Condizioni*
B1	*Sokal-Michener, Abbinamento Semplice*	$S_{xy}^{SM} = \frac{a+d}{p}$	0 1	*nessuno*
B2	*Rogers-Tanimoto*	$S_{xy}^{RT} = \frac{a+d}{p+b+c}$	0 1	*nessuno*
B3	*Jaccard-Tanimoto*	$S_{xy}^{JT} = \frac{a}{a+b+c}$	0 1	$a=0 \rightarrow s=0$
B4	*Gleason-Dice-Sorenson*	$S_{xy}^{GLE} = \frac{2a}{2a+b+c}$	0 1	$a=0 \rightarrow s=0$
B5	*Russel-Rao*	$S_{xy}^{RR} = \frac{a}{p}$	0 1	*nessuno*
B6	*Forbes*	$S^{PERxy} = \frac{pa}{(a+b)(a+c)}$	0 p/a	$den=0 \vee a=0 \rightarrow s=0$
B7	*Simpson*	$S_{xy}^{SIM} = \frac{a}{min\{(a+b),(a+c)\}}$	0 1	$den=0 \vee a=0 \rightarrow s=0$
B8	*Braun-Blanquet*	$S_{xy}^{BB} = \frac{a}{max\{(a+b),(a+c)\}}$	0 1	$a=0 \rightarrow s=0$
B9	*Driver-Kroeber Ochiai coseno*	$S_{xy}^{DK} = \frac{a}{\sqrt{(a+b)(a+c)}}$	0 1	$den=0 \rightarrow s=0$
B10	*Baroni-Urbani-Buser*	$S_{xy}^{BU1} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	0 1	$d=p \rightarrow s=1$
B11	*Kulczynski*	$S_{xy}^{KUL} = \frac{1}{2} \cdot \left[\frac{a}{a+b} + \frac{a}{a+c}\right]$	0 1	$a=0 \rightarrow s=0$

B12	*Sokal-Sneath*	$S_{xy}^{SS1} = \frac{a}{a+2b+2c}$	0 1	$a=0 \rightarrow s=0$
B13	*Sokal-Sneath*	$S_{xy}^{SS2} = \frac{2a+2d}{p+a+d}$	0 1	*nessuno*
B14	*Jaccard*	$S_{xy}^{JA} = \frac{3a}{3a+b+c}$	0 1	$a=0 \rightarrow s=0$
B15	*Fede*	$S_{xy}^{FAI} = \frac{a+0.5 \cdot d}{p}$	0 1	*nessuno*
B16	*Mountford*	$S_{xy}^{MOU} = \frac{2a}{ab+ac+2bc}$	0 2	$den=0 \rightarrow s=a/p$
B17	*Michael*	$S_{xy}^{MIC} = \frac{4 \cdot (ad-bc)}{(a+d)^2+(b+c)^2}$	$+1$ 2	$a=p \vee d=p \rightarrow s=1$ $b+c=0 \rightarrow s=1$
B18	*Rogot-Goldberg*	$S_{xy}^{RG} = \frac{a}{2a+b+c} + \frac{d}{2d+b+c}$	0 1	$a=p \vee d=p \rightarrow s=1$
B19	*Hawkins-Dotson*	$S_{xy}^{HD} = \frac{1}{2} \cdot \left(\frac{a}{a+b+c} + \frac{d}{b+c+d} \right)$	0 1	$a=p \vee d=p \rightarrow s=1$
B20	*Yule*	$S_{xy}^{YU1} = \frac{ad-bc}{ad+bc}$	$+1$ 2	$a=p \vee d=p \vee bc=0 \rightarrow s=1$
B21	*Yule*	$S_{xy}^{YU2} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$+1$ 2	$a=p \vee d=p \vee bc=0 \rightarrow s=1$
B22	*Fossum*	$S_{xy}^{FOS} = \frac{p \cdot (a-0.5)^2}{(a+b)(a+c)}$	0 $\frac{(p-0.5)^2}{p}$	$den=0 \rightarrow s=0$
B23	*Dennis*	$S_{xy}^{DEN} = \frac{ad-bc}{\sqrt{p(a+b)(a+c)}}$	$\frac{\sqrt{p}}{2}$ $\sqrt{p}$	$a=p \vee d=p \rightarrow s=1$ $den=0 \rightarrow s=0$
B24	*Cole*	$S_{xy}^{CO1} = \frac{ad-bc}{(a+c)(c+d)}$	$p-1$ p	$a=p \vee d=p \rightarrow s=1$ $den=0 \rightarrow s=0$
B25	*Cole*	$S_{xy}^{CO2} = \frac{ad-bc}{(a+b)(b+d)}$	$p-1$ p	$a=p \vee d=p \rightarrow s=1$ $den=0 \rightarrow s=0$
B26	*Dispersione*	$S_{xy}^{DIS} = \frac{ad-bc}{p^2}$	$1/4$ $1/2$	$a=p \vee d=p \rightarrow s=1$

B27	*Goodman-Kruskal*	$S_{xy}^{GK} = \frac{2 \cdot \min(a,d) - b - c}{2 \cdot \min(a,d) + b + c}$	+1 2	$a = p \vee d = p \rightarrow s = 1$
B28	*Sokal-Sneath*	$S_{xy}^{SS3} = \frac{1}{4} \cdot \left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right]$	0 1	$a = p \vee d = p \rightarrow s = 1$ $a = 0 \wedge d = 0 \rightarrow s = 0$
B29	*Sokal-Sneath*	$S_{xy}^{SS4} = \frac{a}{\sqrt{(a+b)(a+c)}} \cdot \frac{d}{\sqrt{(b+d)(c+d)}}$	0 1	$a = p \vee d = p \rightarrow s = 1$ $a = 0 \vee d = 0 \rightarrow s = 0$
B30	*Pearson-Heron*	$S_{xy}^{PHI} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	+1 2	$a = p \vee d = p \rightarrow s = 1$ $b = p \vee c = p \rightarrow s = 0$ $den = 0 \rightarrow s = 0$
B31	*Dice-Wallace, Post-Snijders*	$S_{xy}^{DI1} = \frac{a}{(a+b)}$	0 1	$a = 0 \rightarrow s = 0$
B32	*Dice-Wallace, Post-Snijders*	$S_{xy}^{DI2} = \frac{a}{(a+c)}$	0 1	$a = 0 \rightarrow s = 0$
B33	*Sorgenfrei*	$S_{xy}^{SOR} = \frac{a^2}{(a+b)(a+c)}$	0 1	$a = 0 \rightarrow s = 0$
B34	*Cohen*	$S_{xy}^{COE} = \frac{2 \cdot (ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}$	+1 2	$a = p \vee d = p \rightarrow s = 1$ $den = 0 \rightarrow s = 0$
B35	*Peirce*	$S_{xy}^{PE1} = \frac{ad - bc}{(a+b)(c+d)}$	+1 2	$a = p \vee d = p \rightarrow s = 1$ $b = p \vee c = p \rightarrow s = 0$
B36	*Peirce*	$S_{xy}^{PE2} = \frac{ad - bc}{(a+c)(b+d)}$	+1 2	$a = p \vee d = p \rightarrow s = 1$ $b = p \vee c = p \rightarrow s = 0$
B37	*Maxwell-Pilliner*	$S_{xy}^{MP} = \frac{2 \cdot (ad - bc)}{(a+b)(c+d) + (a+c)(b+d)}$	+1 2	$a = p \vee d = p \rightarrow s = 1$ $den = 0 \rightarrow s = 0$
B38	*Harris-Lahey*	$S_{xy}^{HL} = \frac{a \cdot (2d + b + c)}{2 \cdot (a + b + c)} + \frac{d \cdot (2a + b + c)}{2 \cdot (b + c + d)}$	0 p	$a = p \vee d = p \rightarrow s = 1$ $den = 0 \rightarrow s = 0$
B39	*Consonni-Todeschini*	$S_{xy}^{CT1} = \frac{ln(1 + a + d)}{ln(1 + p)}$	0 1	*nessuno*
B40	*Consonni-Todeschini*	$S_{xy}^{CT2} = \frac{ln(1 + p) - ln(1 + b + c)}{ln(1 + p)}$	0 1	*nessuno*
B41	*Consonni-Todeschini*	$S_{xy}^{CT3} = \frac{ln(1 + a)}{ln(1 + p)}$	0 1	*nessuno*

B42	*Consonni-Todeschini*	$S_{xy}^{CT4} = \frac{ln(1+a)}{ln(1+a+b+c)}$	0 1	*nessuno*
B43	*Consonni-Todeschini*	$S_{xy}^{CT5} = \frac{ln\left[\frac{1+ad}{1+bc}\right]}{ln(1+p^2/4)}$	1 2	*nessuno*
B44	*Austin-Colwell*	$S_{xy}^{AC} = \frac{2}{\pi} \cdot arcsin\sqrt{\left(\frac{a+d}{p}\right)}$	0 1	*nessuno*

Esempio 6. Confronto delle distribuzioni delle somiglianze binarie

Le misure di somiglianza per dati binari possono produrre configurazioni diverse delle relazioni all'interno dello stesso dataset, essendo basate su diverse integrazioni degli elementi raccolti nella tabella di frequenza. Come esempio, la Figura 8 raccoglie gli istogrammi che descrivono la distribuzione dei valori di sei somiglianze binarie calcolate sullo stesso dataset che include 1687 campioni (molecole) descritti da 1024 descrittori molecolari binari (impronte) [28]. Come si può vedere, le somiglianze binarie hanno comportamenti diversi, essendo ad esempio la somiglianza Jaccard-Tanimoto (B3) caratterizzata da valori nell'intervallo [0, 0.4], mentre le somiglianze Austin-Colwell (B44), Rogers-Tanimoto (B2) e Sokal-Michener (B1) hanno valori approssimativamente nell'intervallo [0.6, 1].Pertanto, quando si calcola il Scaling Multidimensionale sullo stesso dataset ma con diverse somiglianze binarie, la distribuzione dei campioni nello spazio MDS può cambiare (Figura 9).

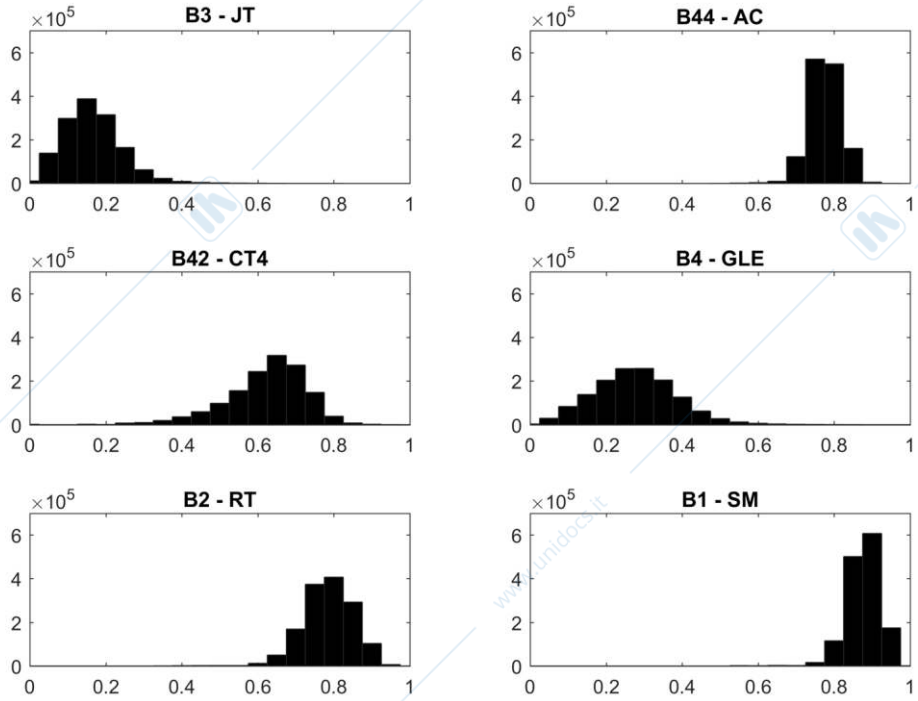

Figura 8. Istogrammi dei valori di sei similarità binarie, calcolati dallo stesso dataset che include 1687 campioni (molecole) descritti da 1024 descrittori molecolari binari (impronte). B3: Jaccard-Tanimoto, B44: Austin-Colwell, B42: Consonni-Todeschini, B4: Gleason-Dice-Sorenson, B2: Rogers-Tanimoto, B1: Sokal-Michener.

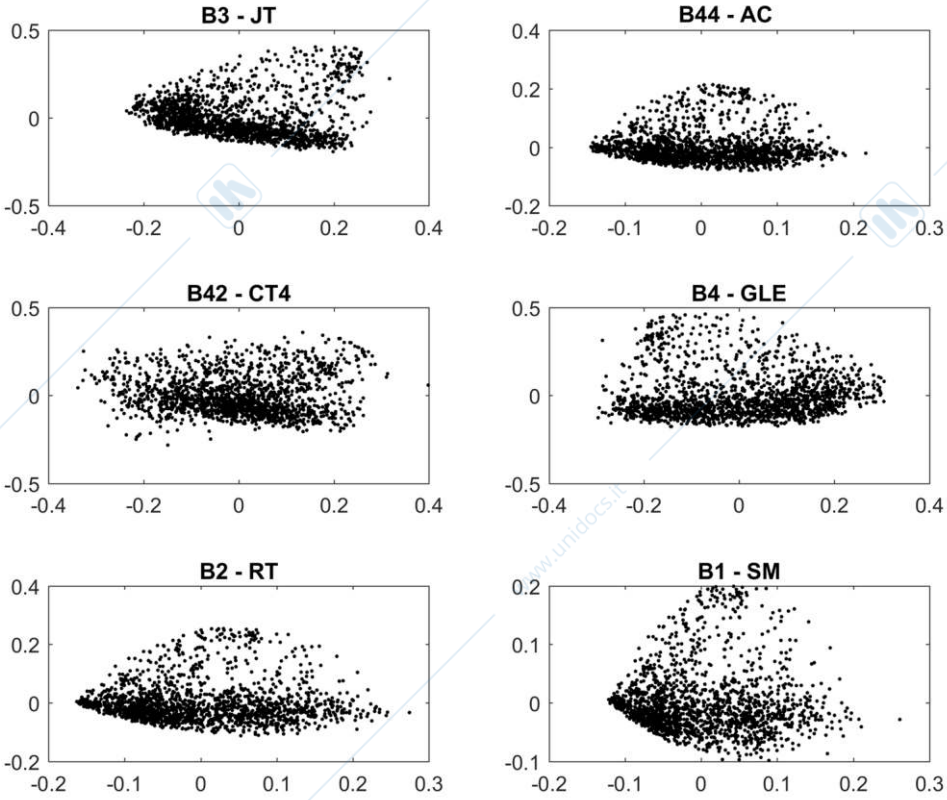

Figura 9. Punteggi di Scaling Multidimensionale ottenuti con sei similarità binarie, calcolati dallo stesso dataset comprendente 1687 campioni (molecole) descritti da 1024 descrittori molecolari binari (impronte). B3: Jaccard-Tanimoto, B44: Austin-Colwell, B42: Consonni-Todeschini, B4: Gleason-Dice-Sorenson, B2: Rogers-Tanimoto, B1: Sokal-Michener.

Mentre i coefficienti di similarità binaria si basano principalmente sul numero a di "presenze comuni" degli attributi e d il numero di "assenze comuni" in x e y, le distanze binarie tengono conto delle voci b e c che forniscono informazioni sulla dissimilarità dei due oggetti. Le misure di distanza binaria più popolari (Tabella 7) sono la distanza di Hamming definita come:

$$D_{xy}^{HAM} = b + c \qquad 0 \le D_{xy}^{HAM} \le p$$

e la distanza di Tanimoto, che è il suo corrispondente normalizzato, definita come:

$$D_{xy}^{T} = \frac{b+c}{p} \qquad 0 \le D_{xy}^{T} \le 1$$

Nella letteratura, sono state proposte e utilizzate anche le loro versioni quadrate radicate (C2 e C4) per applicazioni particolari. Inoltre, altre due misure di distanza binaria sono la distanza non metrica di Watson:

$$D_{xy}^{WAT} = \frac{b+c}{2a+b+c} \qquad 0 \le D_{xy}^{WAT} \le 1$$

e distanza binaria di Soergel:

$$D_{xy}^{SOE} = \frac{b+c}{a+b+c} \qquad 0 \le D_{xy}^{SOE} \le 1$$

Tabella 7. Alcune misure di distanza derivate da variabili binarie.

Eq.	*Distanza*	*Equazione*	*Intervallo*
C1	Distanza di Hamming	$D_{xy}^{HAM} = b+c$	$[0, p]$
C2	Distanza di Hamming quadrata radice	$D_{xy}^{HSR} = \sqrt{b+c}$	$\left[0, \sqrt{p}\right]$
C3	Distanza di Tanimoto	$D_{xy}^{TAN} = \frac{b+c}{p}$	$[0,1]$
C4	Distanza Tanimoto quadrata della radice	$D_{xy}^{TSR} = \sqrt{\frac{b+c}{p}}$	$[0,1]$
C5	Distanza non metrica di Watson	$D_{xy}^{WAT} = \frac{b+c}{2a+b+c}$	$[0,1]$
C6	Distanza binaria di Soergel	$D_{xy}^{SBD} = \frac{b+c}{a+b+c}$	$[0,1]$

Nel confrontare le distanze per variabili binarie e continue a valori reali, è facile vedere che la distanza di Hamming (C1) coincide con la distanza di Manhattan (R2), la distanza di Hamming quadrata (C2) è la distanza euclidea quadrata (R1), la distanza di Tanimoto (C3) coincide con la distanza media di Manhattan e il Tanimoto quadrato (C4) con la distanza euclidea media. Inoltre, la distanza non metrica di Watson (C5) corrisponde alla distanza di Lance-Williams (R11) ed è il complemento del coefficiente di Gleason (B4); la distanza binaria di Soergel (C6) corrisponde alla distanza di Soergel (R13) ed è il complemento del coefficiente di Jaccard/Tanimoto (B3).

Tra le proprietà matematiche dei coefficienti di similarità binaria, particolare attenzione deve essere prestata alla loro metricità, cioè se un coefficiente di similarità può essere trasformato in una distanza metrica. Per definizione, le distanze metriche rispettano l'ineguaglianza triangolare e quelle misure di dissimilarità che non rispettano l'ineguaglianza triangolare sono non metriche o quasi metriche se non è soddisfatta neanche la condizione di simmetria.

Dopo la trasformazione in distanze, è facile vedere che diversi coefficienti di similarità sono non metrici poiché è probabile che due oggetti, x e y, abbiano un valore di distanza maggiore della somma delle loro distanze con un altro oggetto z. Ne consegue che queste misure di similarità non possono essere utilizzate direttamente per proiettare oggetti in uno spazio metrico, a meno che non sia stata applicata una trasformazione adeguata per convertirle in distanze metriche. Inoltre, per ottenere distanze metriche è importante ricordare che qualsiasi trasformazione non induce distanze metriche se la misura di similarità non soddisfa la condizione matematica di simmetria (cioè, $S_{xy} \neq S_{yx}$). Per alcuni coefficienti di similarità binaria, questa condizione non è soddisfatta, per ad esempio, se solo il parametro a o il parametro b, ma non entrambi contemporaneamente, compaiono nella loro definizione: infatti, in questo caso, i valori di b e c scambiano i loro valori. Questo accade, per esempio, per i coefficienti Co1 (B24), Co2 (B25), Di1 (B31) e Di2 (B32). Le proprietà dei coefficienti di similarità binaria sono ulteriormente discusse in un paragrafo successivo di questo lavoro dove viene effettuato un confronto multivariato dei coefficienti di similarità della Tabella 6 su un set di dati simulato.

3.6 Distanze di tipo misto

Nei casi reali, le variabili che descrivono i dati possono essere "miste", cioè possono essere una miscela di valori numerici continui e conteggi (cioè variabili definite in scale di intervallo o di rapporto), classifiche (cioè variabili definite in scale ordinali), attributi categorici e binari (cioè variabili definite in scale nominali).

Pertanto, le distanze di tipo misto, indicate con il simbolo generale DMT, dovrebbero essere utilizzate quando un set di dati contiene variabili di diversi tipi: variabili nominali (n), binarie (b), ordinali (o) e variabili a valore reale (r). In questi casi, per valutare le prossimità di coppie di oggetti, può essere utilizzata la seguente equazione generale:

$$D_{xy}^{MT} = w_n \cdot D_{xy}^{n} + w_b \cdot D_{xy}^{b} + w_o \cdot D_{xy}^{o} + w_r \cdot D_{xy}^{r}$$

dove Dn è il contributo alla distanza calcolato considerando solo variabili nominali, Db il contributo alla distanza calcolato considerando solo variabili binarie, Do il contributo alla distanza calcolato considerando solo variabili ordinali, e Dr il contributo alla distanza calcolato considerando solo variabili a valore reale; wn, wb, wo e wr sono pesi definiti dall'utente per i diversi tipi di contributi alla distanza.

Una misura di similarità generale proposta per affrontare dati di tipo misto è il coefficiente di Gower, che è definito come:

$$S_{xy}^{GOW} = \frac{\sum_{j=1}^{p} s_{xy,j}}{\sum_{j=1}^{p} \delta_{xy,j}}$$

where $s_{xy,j}$ is the similarity of objects x and y calculated for the j-th variable and $\delta_{xy,j}$ è un confronto indice, essendo 1 quando la variabile j può essere utilizzata per confrontare x e y, e 0 altrimenti.

Per le variabili nominali, il contributo di somiglianza è calcolato come:

$$s_{xy,j} = \begin{cases} 1 \text{ se } x = y \\ 0 \text{ altrimenti} \end{cases}$$

Per le variabili binarie, il contributo di similarità e il contatore delle variabili sono calcolati come:

$$s_{xy,j} = \begin{cases} 1 \text{ se } x_j = 1 \wedge s\grave{i} = 1 \\ 0 \text{ altrimenti} \end{cases} \qquad \delta_{xy,j} = \begin{cases} 0 \text{ se } x_j = 0 \wedge s\grave{i} = 0 \\ 1 \text{ altrimenti} \end{cases}$$

Per le variabili a valore reale, il contributo di similarità è calcolato come:

$$s_{xy,j} = 1 - \frac{\left| x_j - s\grave{i} \right|}{U_j - L_j}$$

dove Uj e Lj sono i valori superiore e inferiore della j-esima variabile dei dati.

La distanza di Park è una misura di distanza per dati di tipo misto che richiede che tutte le variabili siano scalate nell'intervallo [0, 1]. È definita come:

- le variabili nominali sono considerate come variabili binarie

- le variabili binarie rimangono inalterate
- le variabili ordinali che assumono valori tra [1, k] sono scalate come $x'_j = x_j / k$
- le variabili a valore reale sono scalate nell'intervallo tra [0, 1].

La distanza del Parco viene quindi calcolata come distanza euclidea media:

$$D_{xy}^{PAR} = \sqrt{\frac{\sum_{j=1}^{p}\left(x_j - s\dot{i}_j\right)^2}{p}}$$

È interessante notare che la distanza di Jaccard-Tanimoto è definita sia per variabili binarie che per variabili a valori reali. Quindi, quando sono presenti solo questi due tipi di variabili nel dataset, la funzione di Jaccard-Tanimoto è una misura adatta delle prossimità tra gli oggetti.

4 Meta-distanze

Il concetto di meta-distanza introduce livelli di misura di similarità/diversità di grado superiore. Questo concetto è stato proposto in [13] per misurare la connessione tra due variabili j e k e chiamato, in quel contesto specifico, Modello di Diffusione del Target Atemporale (ATDM) e successivamente sviluppato ulteriormente in [29].

Le funzioni di distanza delle Tabelle 4 e 5 sono misure di diversità di primo ordine, cioè rilevano le informazioni di base sulle relazioni di diversità tra gli oggetti. Tuttavia, due oggetti possono essere confrontati anche su una scala relativa osservando la loro rispettiva (dis)similarità con tutti gli altri oggetti del set a cui appartengono: più è comparabile la loro similarità con gli oggetti rimanenti, più gli oggetti sono simili.

Sia x e y due oggetti appartenenti a un insieme di n oggetti, allora la meta-distanza $D_{xy}^{(M)}$ è definito come il prodotto tra una distanza primaria D_{xy} (cioè, qualsiasi distanza riportata nelle Tabelle 4 e 5), e un fattore di contrazione α, riferito come la dissimilarità accessoria, che tiene conto delle (dis)similarità di ordine superiore:

$$D_{xy}^{(M)} = \alpha \cdot D_{xy} = e^{-2 \cdot P_{xy}} \cdot D_{xy}$$

dove il termine Pxy che quantifica somiglianze di ordine superiore, è calcolato come:

$$P_{xy} = \frac{1}{n-2} \cdot \sum_{\substack{z \neq x \\ z \neq y}}^{n} \delta(z) \qquad 0 \leq P_{xy} \leq 1$$

$$\delta(z) = \begin{cases} 1 & se \dfrac{1 + \min(D_{xz}, D_{yz})}{1 + \max(D_{xz} D_{yz})} \geq t \\ 0 & altrimenti \end{cases}$$

dove Dxz e Dyz sono le distanze primarie tra qualsiasi z del set, e x e y, rispettivamente; t è una soglia che definisce l'intervallo in cui Dxz e Dyz sono considerati equivalenti (ad esempio, t = 0.95). Il risultato Pxy è la proporzione degli n - 2 oggetti che condividono distanze simili con sia x che y. Di conseguenza, maggiore è questa proporzione, più simili sono x e y a tutti gli altri oggetti, o, in other words, the lower the adjunct dissimilarity α, più simili sono x e y dal punto di vista di degli altri oggetti.

The adjunct dissimilarity α, compreso nell'intervallo [0.1353, 1], funge da parametro di smussamento del primary distance: in fact, if x and y have high proportion of common similar z objects, their α sarà basso e $D_{xy}^{(M)} < D_{xy}$; on the contrary, if no common similar objects are found, α sarà uguale a uno e, quindi, $D_{xy}^{(M)}$Dxy. Se il valore teorico massimo della proporzione di simili comuni

Quando gli oggetti Pxy vengono raggiunti (Pxy = 1), la dissimilarità accessoria è pari a 0,1353 e il valore della distanza primaria è compresso all'86% del suo valore originale. L'introduzione della dissimilarità accessoria influisce sorprendentemente su come viene quantificata la somiglianza tra due oggetti (Figura 10).

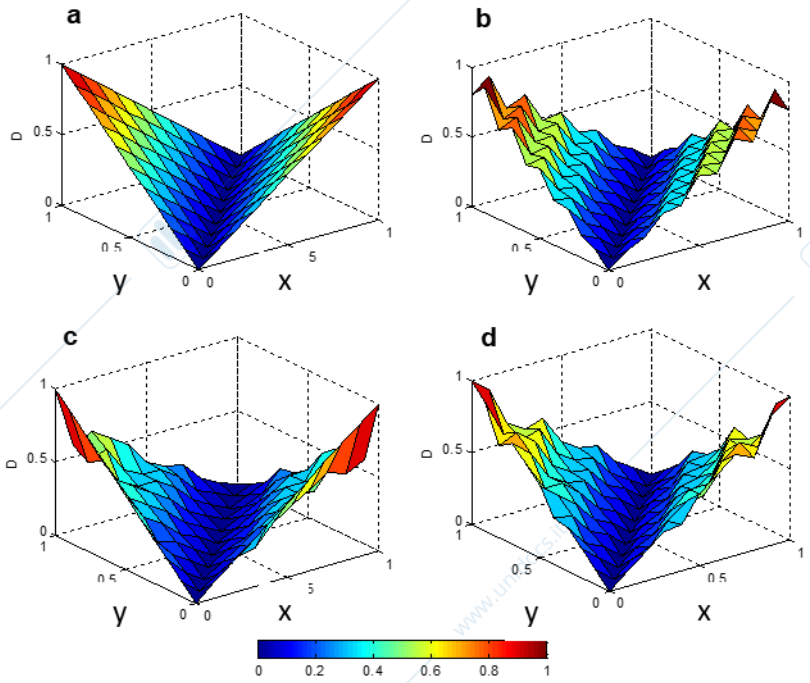

Figura 10 Effetto della dissimilarità degli ausiliari scelti sulla meta-distanza derivata dalla distanza euclidea come contributo principale: (a) Euclidea (solo principale), (b) Euclidea – Euclidea, (c) Euclidea – Lance Williams, (d) Euclidea – Soergel. Due vettori simulati, x e y, sono stati generati da 0 a 1 con un passo di 0,1, e la meta-distanza è stata calcolata in ogni punto della griglia x-y.

Ci sono diverse possibilità per ottenere una meta-distanza, a seconda della scelta della distanza primaria Dxy e della distanza utilizzata per calcolare il fattore di contrazione. In questo studio, ci siamo concentrati sulla misura della meta-distanza derivata dalla distanza Jaccard-Tanimoto DJT come misura di distanza primaria e sulla distanza di Manhattan come distanza del fattore di contrazione.

Matematicamente, questa meta-distanza, chiamata Jaccard-Tanimoto Contratta, è definita come:

$$D_{xy}^{CJT} = D_{xy}^{JT} \cdot e^{-2 \cdot P_{xy}^{MAN}} \qquad 0 \le D_{xy}^{CJT} \le 1$$

$$D_{xy}^{JT} = \left(1 - \frac{\sum_{j=1}^{p} x_j \cdot y_j}{\sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} s_j^2 - \sum_{j=1}^{p} x_j \cdot y_j} \right)^{1/2}$$

$$P_{xy}^{MAN} = \frac{1}{n-2} \cdot \sum_{\substack{z \neq x \\ z \neq y}}^{n} \delta(z) \qquad 0 \le P_{xy}^{MAN} \le 1$$

$$\delta(z) = \begin{cases} 1 & se \ \dfrac{1 + min\left(D_{xz}^{MAN}, D_{yz}^{MAN}\right)}{1 + max\left(D_{xz}^{MAN}, D_{yz}^{MAN}\right)} \ge 0.95 \\ 0 & altrimenti \end{cases}$$

Il valore di 0,95 è stato assunto come soglia al di sotto della quale le due distanze di x e y da z sono considered different, i.e. $\delta(z) = 0$.

5 Distanze tra insiemi

Nel contesto della valutazione di somiglianza/diversità è anche importante il confronto tra due insiemi di oggetti descritti dalle stesse variabili (metriche di collegamento) o tra due insiemi di variabili che descrivono gli stessi oggetti (analisi di Procruste e Misura Canonica della Distanza). Un caso speciale è il confronto di modelli diversi (modelli di regressione o classificazione) per lo stesso insieme di oggetti.

Prima di discutere alcuni approcci specifici per misurare le prossimità tra insiemi, viene brevemente introdotta una misura di distanza generale tra insiemi, vale a dire la distanza di Hausdorff e le sue varianti.

5.1 Distanze di Hausdorff

Quando si valuta la somiglianza tra due sistemi, ciascuno descritto da un insieme discreto finito di elementi, le funzioni adatte per valutare la forza di prossimità tra insiemi di elementi sono la distanza di Hausdorff e le metriche di collegamento utilizzate nei metodi di clustering gerarchico.

Sia A che B due insiemi finiti non vuoti di An e Bn elementi, rispettivamente, ciascun elemento descritto da p variabili. Allora, sia d una misura di distanza e s una misura di somiglianza tra due

elementi nello spazio p-dimensionale. La formula di Hausdorff per misurare la relazione di diversità tra i due insiemi A e B è definita come:

$$D_{\mathrm{AB}}^{HAU} = \max\left\{\sup_{a\in A}\left[\inf_{b\in B}\left(d_{ab}\right)\right], \sup_{b\in B}\left[\inf_{a\in A}\left(d_{ba}\right)\right]\right\}$$

da cui, la misura di similarità corrispondente può essere calcolata come:

$$S_{\mathrm{AB}}^{HAU} = \min\left\{\inf_{a\in A}\left[\sup_{b\in B}\left(s_{ab}\right)\right], \inf_{b\in B}\left[\sup_{a\in A}\left(s_{ba}\right)\right]\right\}$$

dove i simboli d e s si riferiscono rispettivamente alla distanza e alle misure di somiglianza.

Esempi di calcolo delle distanze di Hausdorff sono:

$$D^{HAU}\left([1,7],[3,6]\right) = \max\left[\inf\{2,5\}, \inf\{4,1\}\right] = 2$$

$$D^{HAU}\left(1,[3,6]\right) = \max\left[\inf\{2,5\}\right] = 2$$

$$D^{HAU}\left([4,5],[1,4,5,7]\right) = \max\left[\inf\{3,0,1,3\}, \inf\{4,1,0,2\}\right] = 0$$

L'ultimo caso mostra come per la distanza di Hausdorff $D^{HAU} = 0$ non implica A = B, ma semplicemente che $\mathrm{A} \subseteq \mathrm{B}$. Quindi, in generale, la distanza di Hausdorff è una semi-metrica perché la proprietà di forte riflessività (Axioma 2') non è soddisfatta.

Una distanza di Hausdorff generalizzata è data prendendo la distanza classificata al k-esimo posto piuttosto che il massimo.

Una variante della misura di Hausdorff, chiamata similarità di tipo Hausdorff (Hs) [30], è stata recentemente proposta per tenere conto delle informazioni in tutti gli elementi presenti nei set confrontati:

$$S_{AB}^{Hs} = \frac{\sum_{a\in A}\max_{b\in B}\left[s_{ab}\right] + \sum_{b\in B}\max_{a\in A}\left[s_{ba}\right]}{n_A + n_B}$$

dove abs e bas sono qualsiasi misura di similarità a coppie tra l'elemento p-dimensionale a dell'insieme A e l'elemento p-dimensionale b dell'insieme B. Il termine $\max_{b\in B}\left[s_{ab}\right]$ è la massima somiglianza tra l'elemento a di A e l'insieme B e $\max_{a\in A}\left[s_{ba}\right]$ è la massima somiglianza tra l'elemento b di B e l'insieme A. Pertanto, per ogni elemento di un insieme, la sua massima somiglianza con l'altro insieme è

mantenuta e i contributi massimi di tutti gli elementi dell'insieme vengono quindi mediati sul numero totale di elementi in entrambi gli insiemi.

A differenza della formula di Hausdorff originale che misura il confine esterno degli insiemi, la formula simile a Hausdorff fornisce una somiglianza media complessiva, poiché tiene conto dei contributi di somiglianza di tutti gli elementi dell'insieme invece del contributo di somiglianza minimo (o distanza massima) come fa Hausdorff.

Esempio 7. Somiglianza simile a Hausdorff tra insiemi di elementi

Siano A e B due insiemi finiti non vuoti, ciascuno dei quali include i metaboliti originati da due diverse molecole, rispettivamente Ma e Mb. L'insieme A include due metaboliti (a1 e a2, An uguale a 2), mentre l'insieme B è composto da tre metaboliti (b1, b2, b3, Bn uguale a 3), come mostrato nella Figura 11.

In alcuni casi, come nella valutazione della tossicità delle sostanze chimiche quando entrano negli organismi, può essere utile valutare la somiglianza tra le sostanze chimiche in termini dei loro metaboliti. Pertanto, per stimare la somiglianza tra le molecole Ma e Mb possiamo calcolare la somiglianza di tipo Hausdorff tra A e B (Hs S).

Inizialmente, le somiglianze a coppie tra gli elementi dei due insiemi devono essere $_{AB}$ calcolate. In questo esempio, la struttura chimica di ciascun metabolita è stata rappresentata attraverso impronte digitali binarie e il coefficiente di somiglianza Jaccard-Tanimoto è stato stimato tra ciascuna coppia di metaboliti. I valori numerici sono raccolti nella Tabella 8. Come si può notare nella Figura 11, le sostanze chimiche Ma e Mb condividono due metaboliti comuni e quindi le voci corrispondenti nella Tabella 8 sono associate a un coefficiente di somiglianza Jaccard-Tanimoto binario uguale a 1. Al contrario, i metaboliti a2 e b2 (così come a1 e b3) sono caratterizzati da basse somiglianze strutturali, essendo associati ai coefficienti Jaccard-Tanimoto più bassi (0,30).

Tabella 8. Somiglianze binarie Jaccard-Tanimoto a coppie tra gli elementi degli insiemi A e B. Nell'ultima riga e nelle colonne sono raccolti i valori massimi.

	b1	b2	b3	$\max_{b\in B}\left[s_{ab}\right]$
a1	0.33	1.00	0.30	*1.00*
a2	0.48	0.30	1.00	*1.00*
$\max_{a\in A}\left[s_{ba}\right]$	*0.48*	*1.00*	*1.00*	

La somiglianza di tipo Hausdorff tra A e B può quindi essere calcolata tenendo in considerazione i valori massimi per riga e per colonna della Tabella 8:

$$S_{AB}^{Hs}=\frac{(0.48+1.00+1.00)+(1.00+1.00)}{n_A+n_B}=\frac{4.48}{2+3}=0.896$$

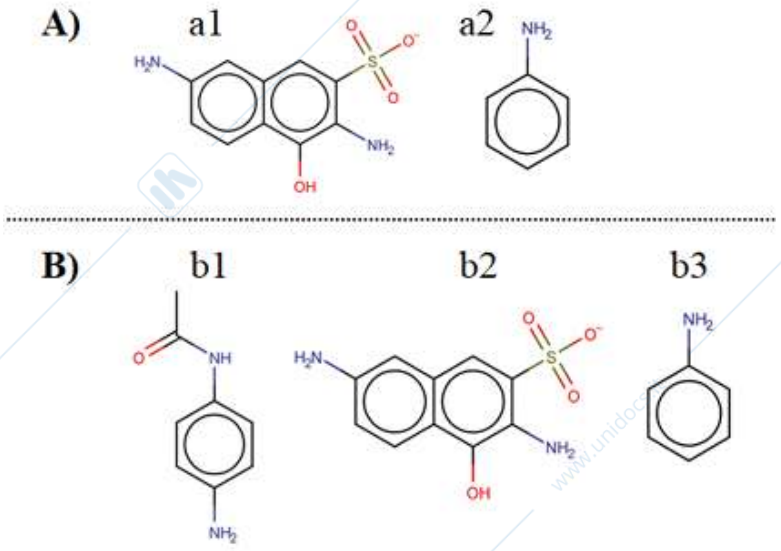

Figura 11 Metaboliti inclusi nei due set A e B.

5.2 Metriche di collegamento

Le metriche di collegamento sono distanze tra due insiemi di oggetti descritti dalle stesse variabili; questi tipi di distanze sono tipicamente utilizzati nell'analisi dei cluster per valutare le prossimità dei cluster.

Sia $A=\{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_M\}$ un insieme di M oggetti e $B=\{\mathbf{b}_1,\mathbf{b}_2,\ldots,\mathbf{b}_N\}$ un insieme di N oggetti, ciascuno descritto dalle stesse p variabili; $\bar{x}_A=\{\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_p\}_A$ e $\bar{x}_B=\{\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_p\}_B$ sono i centroidi p-dimensionali di

i due insiemi (cioè, i vettori dei valori medi delle p variabili che descrivono gli oggetti, calcolati considerando separatamente gli oggetti di ciascun insieme); medA e medB sono le corrispondenti mediane dei due insiemi.

Le metriche di collegamento più comuni sono raccolte nella Tabella 9.

In generale, le metriche di collegamento sono ultrametriche, cioè, l'ineguaglianza ultrametrica è valida (Axiom 4'), che afferma che la distanza tra due oggetti Dxy è minore o uguale alla distanza massima tra ciascuno dei due oggetti e un altro oggetto dell'insieme, cioè,

$$\text{Ax.}4': \quad D_{xy} \leq \max\left\{D_{xz}, D_{zy}\right\}$$

Gli algoritmi del clustering gerarchico agglomerativo sfruttano queste metriche di collegamento per produrre il dendrogramma di un dataset.

Tabella 9. Metriche di collegamento utilizzate nel clustering gerarchico agglomerativo.

Eq.	*Metrica di collegamento*	*Definizione*
L1	*Collegamento Medio*	$D_{\mathrm{AB}}^{AL} = \left(\sum_{a=1}^{M}\sum_{b=1}^{N} d_{ab}\right) / M \cdot N$
L2	*Collegamento Singolo*	$D_{\mathrm{AB}}^{CS} = \min_{\mathrm{a,b}}\left(d_{ab}\right)$
L3	*Collegamento Completo*	$D_{\mathrm{AB}}^{CL} = \max_{\mathrm{a,b}}\left(d_{ab}\right)$
L4	*Collegamento del centroide*	$D_{\mathrm{AB}}^{CEN} = \left(\mathbf{c}_{\mathrm{A}} - \mathbf{c}_{\mathrm{B}}\right)^2$
L5	*Collegamento mediano*	$D_{\mathrm{AB}}^{MED} = \left(\mathbf{med}_{\mathrm{A}} - \mathbf{med}_{\mathrm{B}}\right)^2$
L6	*Collegamento del reparto*	$D_{\mathrm{AB}}^{WL} = \sqrt{\frac{M \cdot N}{M + N}} \cdot \left(\mathbf{c}_{\mathrm{A}} - \mathbf{c}_{\mathrm{B}}\right)^2$

5.3 Analisi di Procruste

L'analisi di Procruste è un metodo statistico per confrontare due set di dati composti dagli stessi oggetti ma descritti con diversi insiemi di variabili. I due set di dati potrebbero essere, ad esempio, gli insiemi di variabili di due diversi modelli di classificazione o regressione ottenuti dallo stesso insieme di oggetti.

L'analisi di Procruste determina una trasformazione lineare, basata su traslazione, riflessione, rotazione ortogonale e scalatura, dei punti nel primo dataset per conformarli al meglio ai punti nel secondo dataset [31,1]. Il criterio di bontà di adattamento di Procruste misura in qualche modo la dissimilarità tra i due dataset, essendo la somma delle differenze quadrate tra i punti dopo traslazione, dilatazione e rotazione di un dataset rispetto all'altro; è uguale a 0 se i due dataset coincidono, mentre è uguale a 1 se le strutture dei dati sono completamente dissimili.

5.4 Misura canonica della distanza

La Misura Canonica della Distanza (CMD) è una funzione di dissimilarità proposta per confrontare due dataset con gli stessi oggetti ma due diversi insiemi di variabili come per l'analisi di Procruste.

Siano A e B i due diversi dataset. Il modo più semplice per misurare la distanza tra questi due dataset ignora i valori delle variabili effettive e consiste semplicemente nel calcolare il numero di variabili diverse nei due dataset. Questa funzione è la distanza di Hamming (Tabella 7, C1), definita per due insiemi A e B come:

$$D_{\mathrm{AB}}^{HAM} = b + c$$

dove b è il numero di variabili in A ma non in B, e c il numero di variabili presenti in B ma non in A. La distanza di Hamming ha solitamente un bias verso l'alto poiché sovrastima la distanza reale tra due insiemi di variabili, a causa del fatto che la correlazione tra variabili non è presa in considerazione.

La Misura Canonica di Distanza [32] supera questo svantaggio ed è definita come:

$$D_{\mathrm{AB}}^{CMD} = p_{\mathrm{A}} + p_{\mathrm{B}} - 2 \cdot \sum_{j=1}^{M} \sqrt{\lambda_j} \qquad 0 \le D_{\mathrm{AB}}^{CMD} \le \left(p_{\mathrm{A}} + p_{\mathrm{B}}\right)$$

where p_{A} and p_{B} are the number of variables in set A and B, respectively; λ sono i valori propri di matrice di cross-correlazione simmetrica e M è il numero di autovalori non nulli.

La matrice di cross-correlazione contiene i coefficienti di correlazione a coppie tra le variabili del two sets; it is an unsymmetrical matrix $\mathbf{C_{AB}}$ of size $(p_{\mathrm{A}} \times p_{\mathrm{B}})$ or $\mathbf{C_{BA}}$ of size $(p_{\mathrm{B}} \times$ pA). La matrice di cross-correlazione simmetrica è derivata dal seguente prodotto interno:

$$\mathbf{Q_A = C_{AB} \times C_{BA}} \qquad \text{oppure} \qquad \mathbf{Q_B = C_{BA} \times C_{AB}}$$

where $\mathbf{Q}_A$ and $\mathbf{Q}_B$ are two different square symmetrical matrices, one of size $p_A \times$ pA e l'altro di size $p_B \times$ pB. Sebbene queste matrici simmetriche siano diverse, i loro valori propri non nulli M coincidono, M essendo il rango minimo tra QA e QB.

La Misura Canonica di Correlazione (CMC) è stata anche derivata dai valori propri non nulli λ delle matrici di cross-correlazione simmetriche come segue:

$$\rho_{AB}^{CMC} \equiv CMC_{AB} = \frac{\sum_{j=1}^{M} \sqrt{\lambda_j}}{\sqrt{p_A \cdot p_B}} \qquad 0 \le \rho_{AB}^{CMC} \le 1$$

dove il numeratore misura la varianza comune inter-set e il denominatore è il suo valore massimo teorico. Questo indice è correlato alla struttura correlazionale multidimensionale tra due insiemi di variabili. Se non esiste correlazione tra nessuna coppia di variabili dei due insiemi, allora CMC= 0 e l'indice CMD si riduce alla distanza di Hamming.

La funzione CMD soddisfa i primi tre assiomi principali per una misura di distanza; tuttavia, l'ineguaglianza triangolare non è sempre valida, quindi la misura canonica di distanza tra insiemi è una distanza non metrica.

Esempio 8. Misure di somiglianza/diversità tra insiemi.

Un esempio di calcolo delle misure di prossimità tra insiemi è presentato di seguito. Sono stati generati diversi set di dati, ciascuno ottenuto combinando in modi diversi quattro variabili. La Tabella 10 mostra le correlazioni tra le variabili a coppie, mentre le misure CMD e CMC sono fornite nella Tabella 11, insieme alla distanza di Hamming e al criterio di bontà di adattamento di Procrustes.

Tabella 10. Correlazioni a coppie tra le variabili x1, x2, x3 e x4.

	x_1	x_2	x_3	x_4
x_1	1	0.979	0.061	0.475
x_2	0.979	1	0.194	0.593

x_3	0.061	0.194	1	0.240
x_4	0.475	0.593	0.240	1

Tabella 11. Distanza di Hamming DHAM, distanza di Procruste DPR, indici CMD e CMC per diversi set di variabili; le correlazioni tra coppie di variabili sono raccolte nella Tabella 11. pA, pB, b e c sono i termini definiti nel testo.

ID	*Set A*	*Set B*	p_A	p_B	*b*	*c*	D^{HAM}	D^{PR}	*CMD*	*CMC*
1	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	4	4	0	0	0	0	0	1
2	$\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	3	3	1	1	2	0.004	0.028	0.995
3	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4$	4	3	1	0	1	0.017	0.150	0.989
4	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	4	3	1	0	1	0.019	0.176	0.985
5	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	4	3	1	0	1	0.338	0.591	0.925
6	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4$	3	3	1	1	2	0.412	0.651	0.892
7	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	3	3	1	1	2	0.412	0.681	0.887
8	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4$	4	3	1	0	1	0.109	0.819	0.892
9	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_2, \mathbf{x}_3$	4	2	2	0	2	0.377	0.927	0.897
10	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_1, \mathbf{x}_3$	4	2	2	0	2	0.431	0.993	0.885
11	$\mathbf{x}_1$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4$	1	3	0	2	2	0.491	1.045	0.853
12	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_1, \mathbf{x}_4$	4	2	2	0	2	0.144	1.100	0.866
13	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_2, \mathbf{x}_4$	4	2	2	0	2	0.147	1.127	0.861
14	$\mathbf{x}_1$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	1	3	0	2	2	0.309	1.199	0.809
15	$\mathbf{x}_3$	$\mathbf{x}_4$	1	1	1	1	2	0.942	1.520	0.240
16	$\mathbf{x}_1$	$\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	1	3	1	3	4	0.662	1.821	0.629
17	$\mathbf{x}_1$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	1	4	0	3	3	0.559	2.042	0.740
18	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4$	$\mathbf{x}_3, \mathbf{x}_4$	2	2	2	1	3	0.384	2.272	0.432
19	$\mathbf{x}_1, \mathbf{x}_2$	$\mathbf{x}_3, \mathbf{x}_4$	2	2	2	2	4	0.740	2.291	0.427
20	$\mathbf{x}_3$	$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4$	1	3	1	1	4	0.956	3.371	0.182

Dai risultati della Tabella 11, la prima considerazione è che l'indice CMD rimuove la forte degenerazione della distanza di Hamming, consentendo una migliore distinzione dei diversi casi. È anche evidente che le conclusioni derivate dalla distanza di Hamming sono piuttosto diverse dalle conclusioni basate sull'indice CMD. Consideriamo, ad esempio, il caso 2 che si riferisce al confronto di due set di dati che hanno due variabili comuni (x3 e x4) e la terza variabile diversa. La distanza di Hamming è uguale a 2, il che significa che c'è una certa differenza tra

i due set di dati; al contrario, l'indice CMD è molto vicino a zero, il che significa che i due set sono in realtà gli stessi. Questa è una conseguenza della grande correlazione (cioè 0,979) tra le variabili x1 e x2. Considerazioni simili possono essere fatte su tutti gli altri casi mostrati nella Tabella 11.

Un'altra misura dai set di dati può essere definita dall'indice di correlazione multivariata K [33,34]. Infatti, semplicemente riguardando la correlazione globale in una matrice di dati X (n x p), l'indice di correlazione multivariata K è stato definito come:

$$K = \frac{\sum_{k=1}^{p} \left| \frac{\lambda_k - 1}{p} \right|}{\frac{2 \cdot (p-1)}{p}} \qquad 0 \le K \le 1$$

where λ sono gli autovalori della matrice di correlazione C, calcolata a partire da una matrice di dati X, e p è il rango minimo tra n e p. Questa espressione è una generalizzazione della nota correlazione bivariata, che, in termini di autovalori, assume la forma:

$$K = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2} \ .$$

Da questo indice, la distanza di correlazione globale tra due set di dati X e Y in termini di correlazione può essere naturalmente definita come:

$$D_{XY}^{Kcor} = \left| K_{\mathbf{X}} - K_{\mathbf{Y}} \right| \qquad 0 \le D_{XY}^{Kcor} \le 1$$

dove le due matrici di dati possono avere dimensioni diverse, sia per righe che per colonne.

6 Misure di distanza sui grafi

Un grafo è solitamente denotato come G = (V, E), dove V è un insieme di vertici e E è un insieme di elementi che rappresentano la relazione binaria tra coppie di vertici; le coppie di vertici non ordinate sono chiamate archi.

Diversi sistemi possono essere rappresentati da grafi, come ad esempio reti sociali, reti di collaborazione, reti di comunicazione, reti bibliometriche e così via. Nel campo della chimica, i grafi sono utilizzati per rappresentare le molecole e sono specificamente chiamati grafi molecolari, dove

I vertici e i bordi sono interpretati come atomi e legami chimici. Un grafo molecolare rappresenta la connettività degli atomi in una molecola indipendentemente dai parametri metrici come le distanze interatomiche di equilibrio tra i nuclei, gli angoli di legame e gli angoli di torsione, rappresentando la geometria molecolare 3D.

Le distanze tra vertici e bordi sono principalmente calcolate in termini di distanze topologiche e di deviazione.

La distanza topologica D_{xy}^{TOP} è il numero di spigoli lungo il percorso più breve tra i vertici vx e vy, cioè, la lunghezza del geodetico tra vx e vy.

La distanza di deviazione Δ_{xy}^{DET} è esattamente l'"opposto" della definizione della distanza topologica che è la lunghezza del percorso più lungo tra i vertici vx e vy, cioè è il numero massimo di spigoli che separano i due vertici.

Si può notare che le distanze topologiche e di deviazione coincidono per i grafi aciclici, poiché esiste solo un percorso che collega qualsiasi coppia di vertici, mentre possono differire quando è presente almeno un ciclo nel grafo.

7 Confronto multivariato delle distanze a valori reali

Per eseguire un confronto multivariato tra le distanze a valori reali, sono stati considerati otto set di dati di riferimento. L'elenco di questi set di dati, insieme al numero di oggetti e variabili, è fornito nella Tabella 12. Per ciascun set di dati, la partizione degli oggetti in diverse classi è stata utilizzata per valutare gli effetti delle misure di distanza sulla classificazione basata sulla similarità [35,36,37,38,39,40,41,42].

Tabella 12. Elenco dei set di dati utilizzati per confrontare le distanze per dati a valori reali.

ID	*Set di dati*	*Oggetti*	*Variabili*	*Classi*
1	Iris	150	4	3
2	Vini	178	13	3
3	Perpot	100	2	2
4	Sulfa	50	7	2

5	Tiofene	24	3	3
6	Itaoils	572	8	9
7	Sangue	784	4	2
8	Diabete	768	8	2

Le seguenti 18 misure di distanza sono state considerate per questa analisi:

- 9 distanze illimitate (Tabella 4), vale a dire Euclidea (EUC, R1), Manhattan (MAN, R2), Lagrange (LAG, R3), Bhattacharyya (BHA, R5), Mahalanobis (MAH, R6) e le 4 diverse distanze Mahalanobis localmente centrate (R7); queste ultime sono le Mahalanobis localmente centrate asimmetriche
- centrate in x (MU), le Mahalanobis localmente centrate asimmetriche centrate in y (ML), le Mahalanobis localmente centrate simmetriche basate sulla media aritmetica (MSA), le Mahalanobis localmente centrate simmetriche basate sulla media geometrica (MSG);
-

8 distanze limitate (Tabella 5), vale a dire Canberra (CAN, R9), Clark (CLA, R10), Wave-Edge (WE, R11), Lance-Williams (LW, R12), Soergel (SOE, R13), Jaccard-Tanimoto (JT, R16), Correlazione (COR, R18), Coseno (CD, R19);

1 meta-distanza, ovvero la distanza Jaccard-Tanimoto contratta (CJT), derivata dalla distanza Jaccard-Tanimoto come distanza primaria e dalla distanza di Manhattan come fattore di contrazione della meta-similarità.

7.1 Confronto delle misure di distanza a valori reali nell'analisi non supervisionata

Questo studio è stato intrapreso con l'obiettivo di indagare come le diverse misure di distanza influenzino le relazioni reciproche tra gli oggetti di un dataset e, pertanto, come la loro visualizzazione grafica e i risultati dell'analisi non supervisionata possano cambiare di conseguenza.

Un esempio visivo di come le 18 distanze considerate inducano diverse relazioni di somiglianza/diversità tra gli oggetti e quindi definiscano diverse geometrie dei dati è mostrato nella Figura 12.

per il dataset Iris. I grafici della Figura 12 sono le proiezioni dei dati in uno spazio bidimensionale ottenute mediante la tecnica di Scaling Multidimensionale, che è un metodo multivariato adatto a tenere conto delle relazioni reciproche delle distanze degli oggetti, riproducendo la struttura dei dati codificata nella matrice delle distanze (similarità) in uno spazio a bassa dimensione.

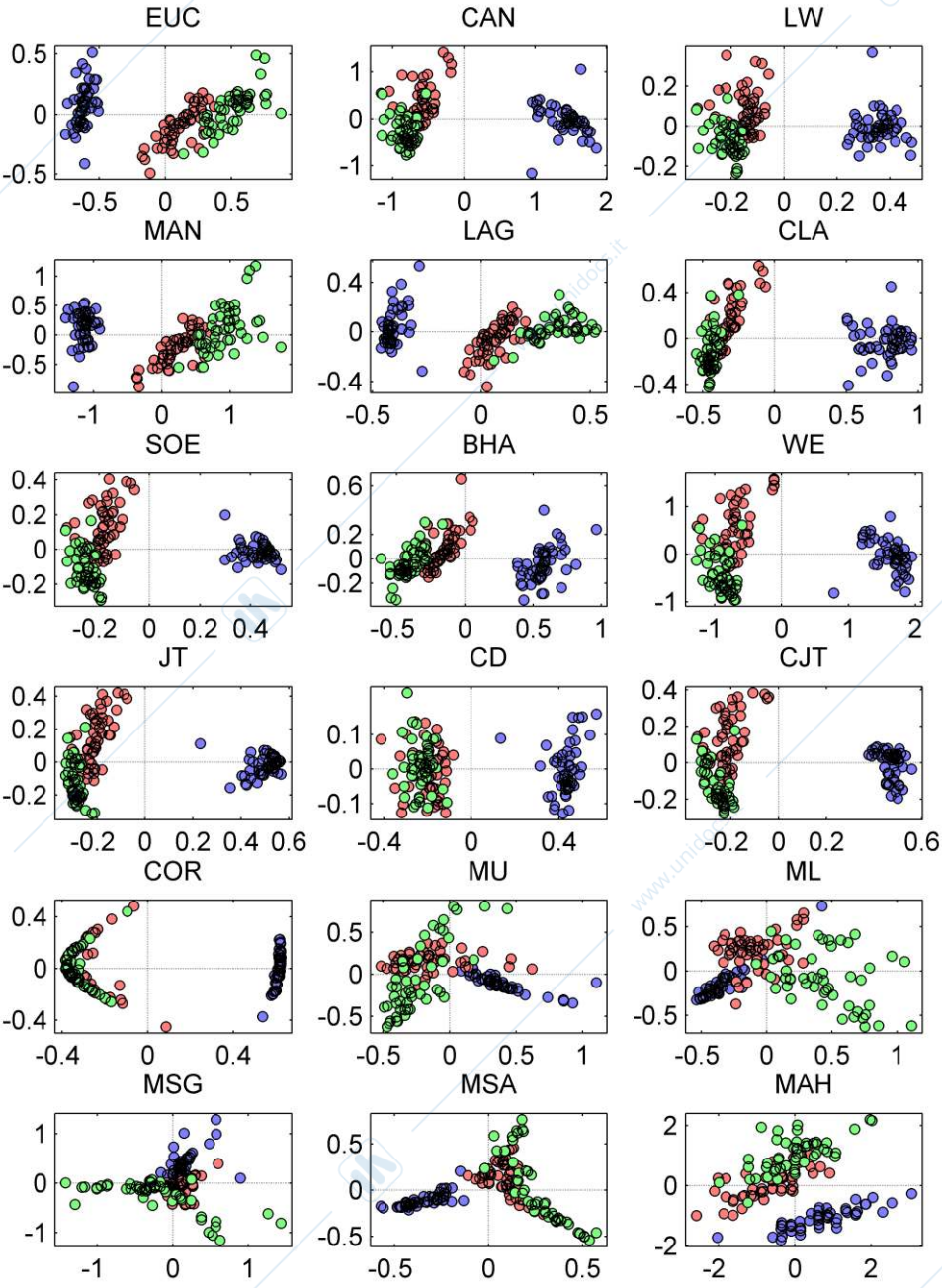

Figura 12 Proiezioni dei 150 oggetti Iris mediante Scaling Multidimensionale basato su 18 misure di distanza. I diversi colori rappresentano le classi.

Le modifiche nella distribuzione delle tre classi del dataset Iris consentono un facile confronto degli effetti delle distanze sulle relazioni tra gli oggetti.

A prima vista, la maggior parte delle distanze rivela una separazione tra la classe blu rispetto alle altre. D'altra parte, questo comportamento non è così evidente per le 4 distanze – ovvero, MU, ML, MSG, MSA – derivate dalla distanza di Mahalanobis Localmente Centrata (LCM), che offrono una visione diversa della distribuzione delle classi. Rispetto alle classi verde e rossa, le distanze forniscono diversi gradi di separazione; ad esempio, queste due classi si sovrappongono quando si utilizzano CD e COR, mentre migliori separazioni visive si ottengono tramite EUC, LAG e CJT. Infine, ogni distanza offre una percezione diversa della presenza di outlier; ad esempio, quasi tutte le distanze rilevano uno o più outlier nella classe blu, mentre CJT fornisce una clusterizzazione compatta di questa classe.

Per ogni dataset in analisi, le dissimilarità a coppie sono state quindi calcolate tra tutte le possibili pairs of objects (i.e., $n \cdot (n-1)/2$, n essendo il numero totale di oggetti del dataset) utilizzando tutte le 18 funzioni di distanza considerate, una alla volta. Alla fine di questo calcolo, le distanze a coppie were collected into a data matrix of dimension $n \cdot (n-1)/2$ × 18, dove le righe rappresentano le coppie di oggetti e le colonne le misure di distanza considerate. L'Analisi delle Componenti Principali è stata quindi applicata a questa matrice di dati per indagare le relazioni tra le diverse funzioni di distanza.

I grafici di caricamento dei primi quattro PC sono riportati nelle Figure 13 – 20; i primi quattro PC spiegano sempre una varianza totale superiore al 90%. Per rendere i grafici più leggibili, sono stati adottati i seguenti colori: verde per le 5 distanze di tipo Mahalanobis (MAH: Mahalanobis classico, ML: funzione Mahalanobis localmente centrata in x, MU: funzione Mahalanobis localmente centrata in y, MSA: media aritmetica simmetrica localmente centrata di Mahalanobis, MSG: media geometrica simmetrica localmente centrata di Mahalanobis); nero per le 2 distanze limitate basate sulla correlazione (COR: distanza di correlazione) e (CD: distanza coseno); blu per altre 4 distanze illimitate (EUC: Euclidea, MAN: Manhattan, LAG: Lagrange, BHA: Bhattacharyya);

rosso per altre 6 distanze limitate (CAN: Canberra, LW: Lance-Williams, WE: Wave-Edge, CLA:

Clark, SOE: Soergel, JT: Jaccard-Tanimoto); pimk: 1 meta-distanza (CJT: Jaccard-Tanimoto contratta).

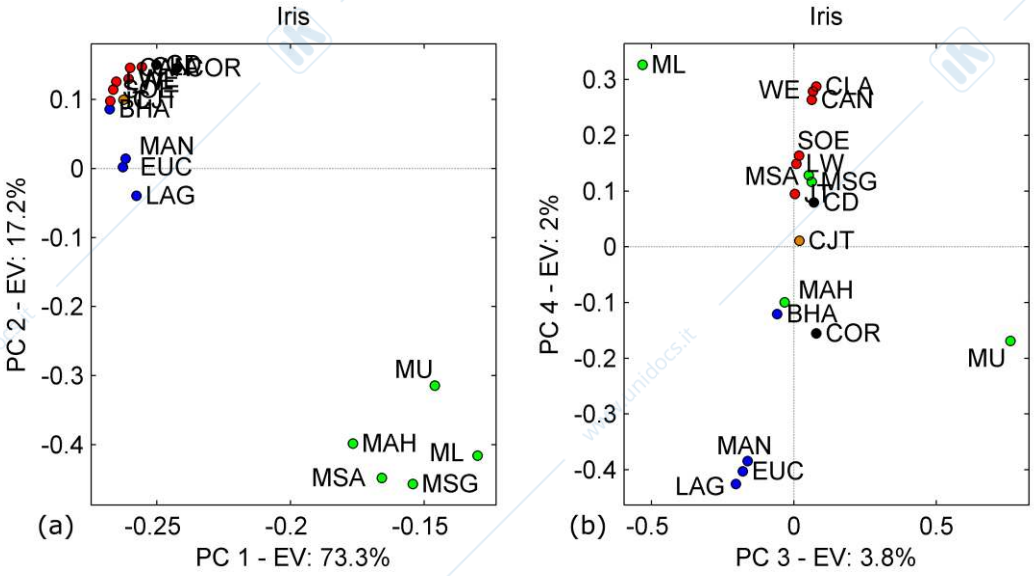

Figura 13 PCA delle distanze a coppie tra gli oggetti del dataset Iris ottenute dalle diverse funzioni di distanza. (a) Grafico di caricamento PC1-PC2. (b) Grafico di caricamento PC3-PC4.

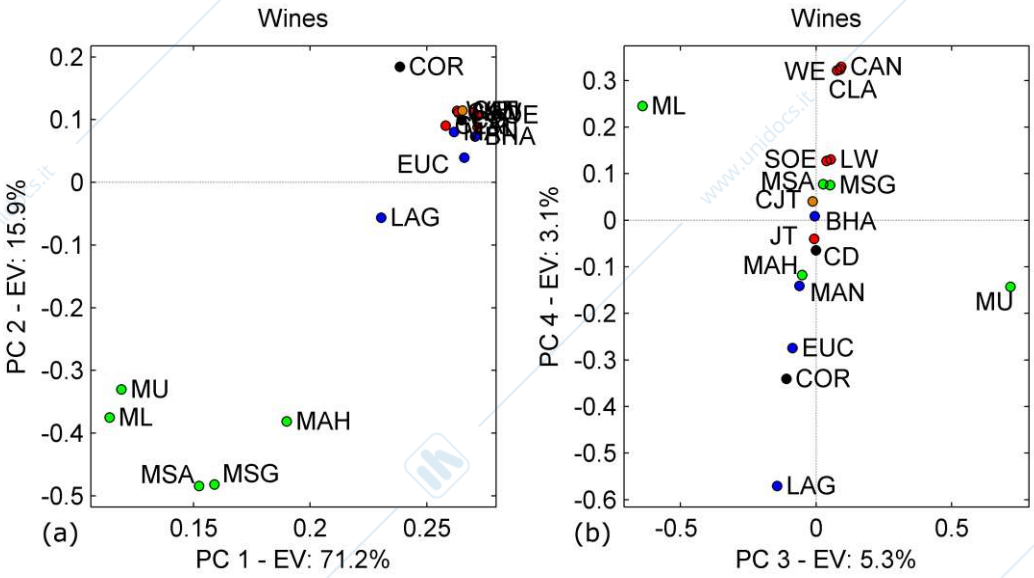

Figura 14 PCA delle distanze a coppie tra gli oggetti del dataset Vini ottenute dalle diverse funzioni di distanza. (a) Grafico di caricamento PC1-PC2. (b) Grafico di caricamento PC3-PC4.

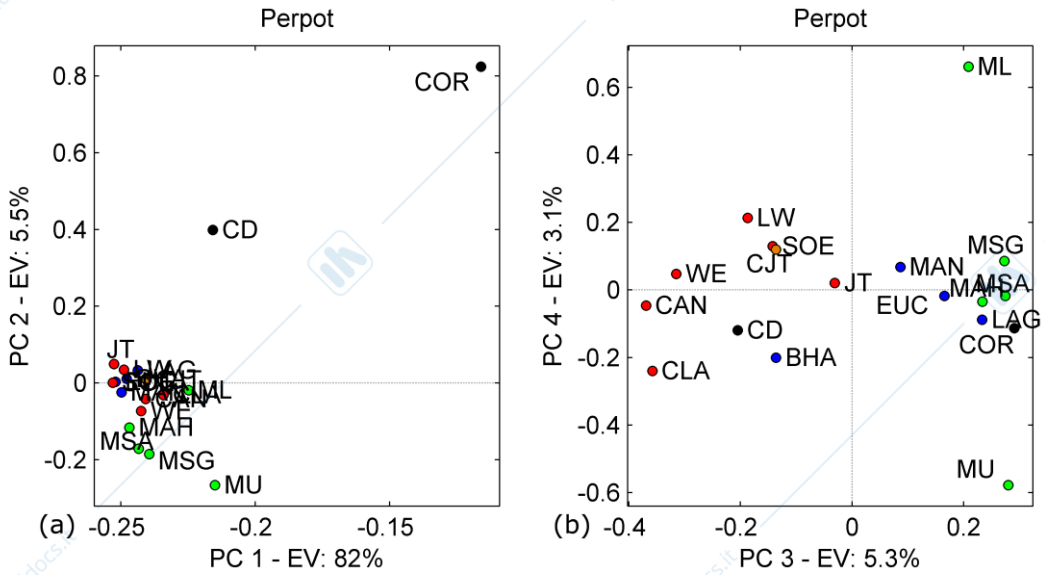

Figura 15 PCA delle distanze a coppie tra gli oggetti del dataset Perpot ottenute dalle diverse funzioni di distanza. (a) Grafico di caricamento PC1-PC2. (b) Grafico di caricamento PC3-PC4.

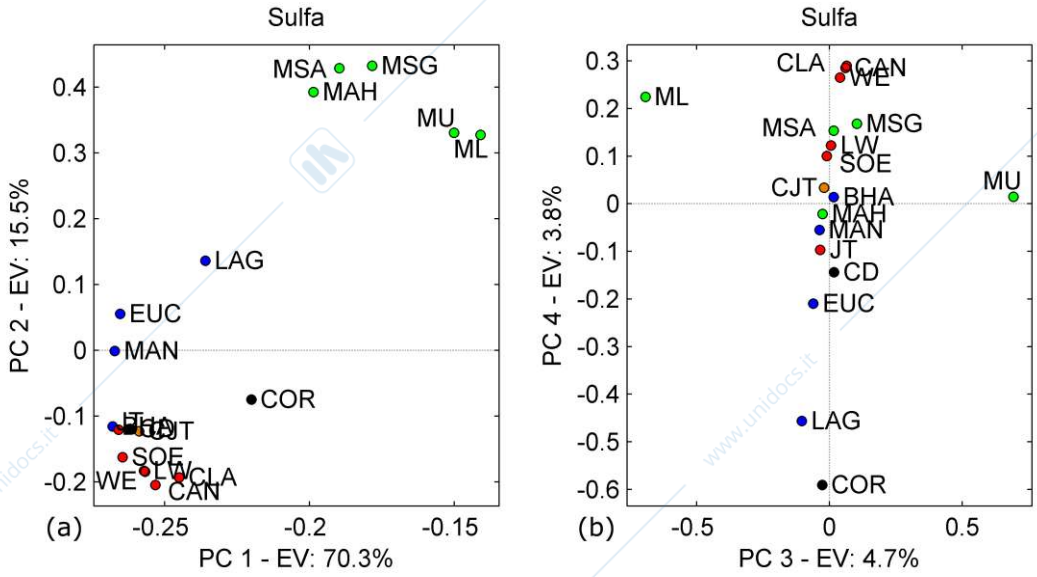

Figura 16 PCA delle distanze a coppie tra gli oggetti del dataset Sulfa ottenute dalle diverse funzioni di distanza. (a) Grafico di caricamento PC1-PC2. (b) Grafico di caricamento PC3-PC4.

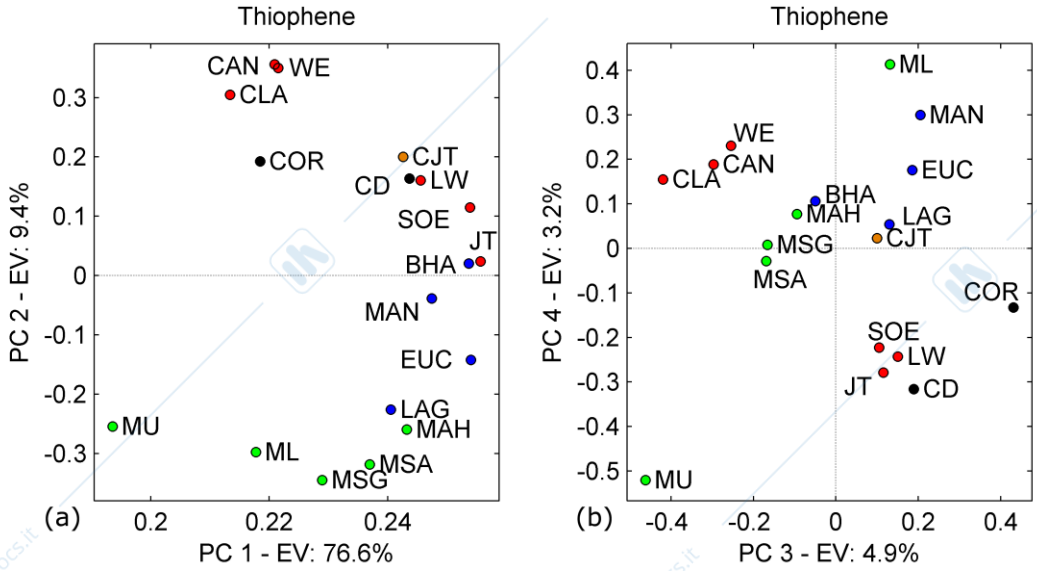

Figura 17 PCA delle distanze a coppie tra gli oggetti del dataset Thiophene ottenute mediante le diverse funzioni di distanza. (a) Grafico di caricamento PC1-PC2. (b) Grafico di caricamento PC3-PC4.

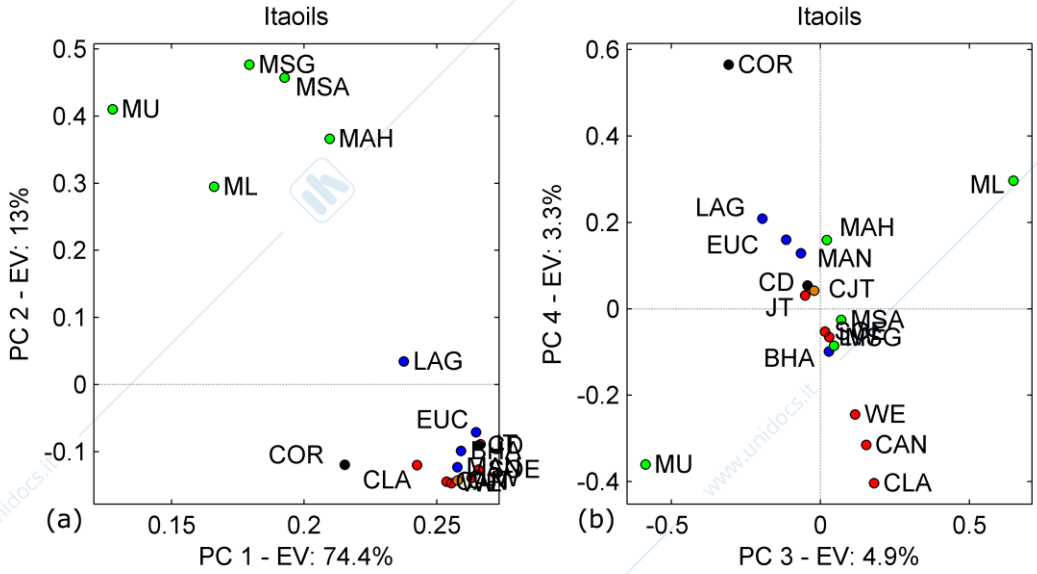

Figura 18 PCA delle distanze a coppie tra gli oggetti del dataset Itaoils ottenute mediante le diverse funzioni di distanza. (a) Grafico di caricamento PC1-PC2. (b) Grafico di caricamento PC3-PC4.

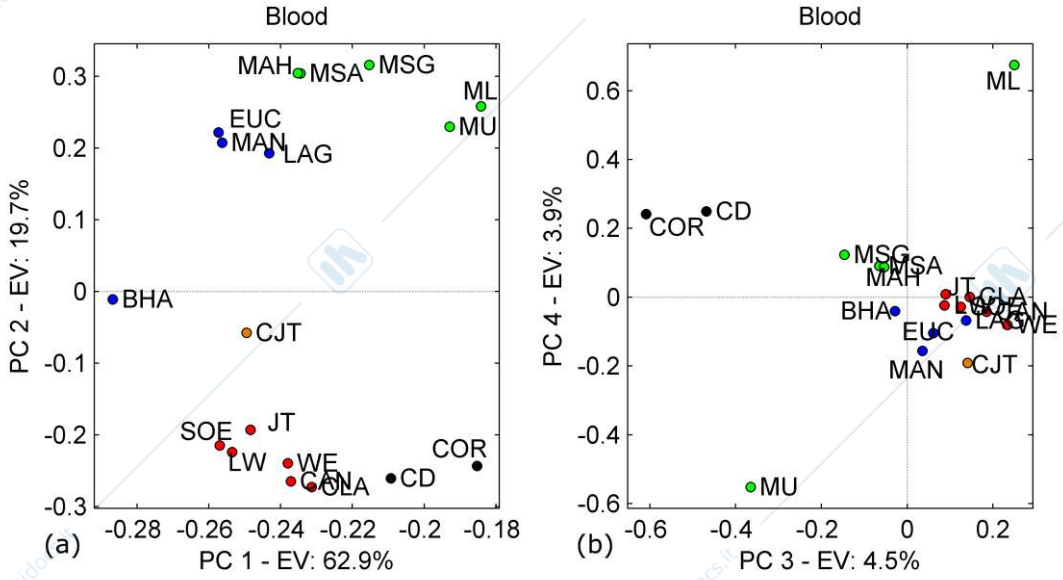

Figura 19 PCA delle distanze a coppie tra gli oggetti del dataset Blood ottenute mediante le diverse funzioni di distanza. (a) Grafico di caricamento PC1-PC2. (b) Grafico di caricamento PC3-PC4.

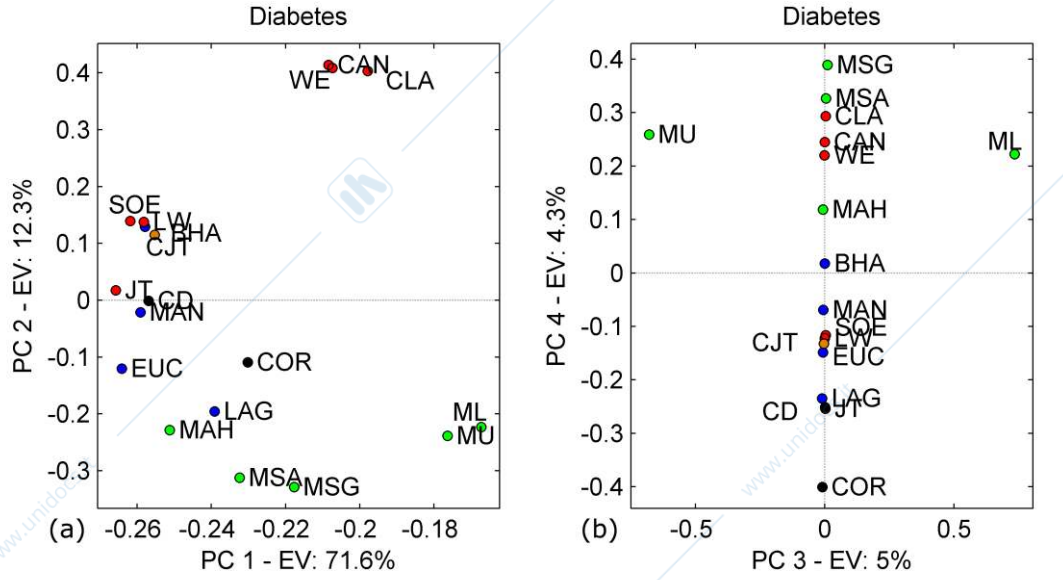

Figura 20 PCA delle distanze a coppie tra gli oggetti del dataset Diabetes ottenute mediante le diverse funzioni di distanza. (a) Grafico di caricamento PC1-PC2. (b) Grafico di caricamento PC3-PC4.

In primo luogo, si può notare che nella maggior parte dei casi le cinque distanze di tipo Mahalanobis si comportano in modo diverso rispetto alle altre distanze. In particolare, questa separazione è evidente per i dataset Iris (Figura 13), Itaoils (Figura 18), Sulfa (Figura 16), Wines (Figura 14) e Blood (Figura 19). Inoltre, la terza o la quarta componente evidenzia il comportamento opposto delle due funzioni Mahalanobis asimmetriche localmente centrate MU e ML, mentre le due diverse procedure di simmetrizzazione di LCM.

le distanze, cioè la media aritmetica (MSA) e la media geometrica (MSG), non sono significativamente distinguibili in tutte le PC per tutti i dataset.

Le distanze euclidea (EUC) e di Manhattan (MAN) sono spesso molto simili e, in diversi casi, non lontane dalla distanza di Lagrange (LAG). Per i set di dati Sangue (Figura 19), Diabete (Figura 20) e Tiofene (Figura 17), sono più vicine al gruppo delle distanze di tipo Mahalanobis rispetto alle distanze limitate. In particolare, la distanza di Lagrange (LAG) sembra essere simile alla distanza di Mahalanobis (MAH) nei set di dati Diabete e Tiofene ed è, in generale, la più simile al gruppo delle distanze di tipo Mahalanobis. La distanza di Bhattacharyya (BHA) sembra essere diversa dalle altre distanze illimitate (vedi, ad esempio, Diabete, Sangue) ed è spesso simile alla distanza di Jaccard-Tanimoto (JT).

In alcuni casi, la distanza di correlazione (COR) appare separata dalle altre distanze; in misura minore, lo stesso comportamento è mostrato anche dalla distanza coseno (CD). Questo può essere chiaramente notato nei dataset Perpot (Figura 15) e Blood (Figura 19).

Oltre alla correlazione e alle distanze coseno, le altre sei distanze limitate, ovvero Lance-Williams (LW), Canberra (CAN), Wave-Edge (WE), Soergel (SOE), Jaccard-Tanimoto (JT) e Clark (CLA), appaiono molto spesso nella stessa regione dello spazio PC, specialmente nei primi due componenti. Queste distanze possono essere ulteriormente suddivise in due sottogruppi, il primo costituito dalle distanze Wave-Edge (WE), Canberra (CAN) e Clark (CLA) e il secondo da Lance-Williams (LW), Soergel (SOE) e Jaccard-Tanimoto (JT).Ad esempio, evidenze di questi due sottogruppi possono essere trovate nei primi due PC dei dataset Blood (Figura 19a), Diabetes (Figura 20a) e Thiophene (Figura 17a), e nei terzi-quarti PC dei dataset Iris (Figura 13b), Thiophene (Figura 17b), Sulfa (Figura 16) e Itaoils (Figura 18b).

Inoltre, è notevole che la distanza Jaccard-Tanimoto Contratta (JCT) di solito non è molto lontana dalla distanza Jaccard-Tanimoto (JT) come previsto; tuttavia, appare chiaramente distinguibile da JT in tutti gli 8 dataset. Queste relazioni possono essere facilmente notate nei primi due PC per i dati.

imposta Blood (Figura 19a), Thiophene (Figura 17a) e Diabetes (Figura 20a) e nei terzi e quarti PC per i set di dati Thiophene (Figura 17b), Sulfa (Figura 16b) e Perpot (Figura 15b).

7.2 Effetti delle misure di distanza sulla classificazione basata sulla similarità

I metodi di classificazione sono tecniche multivariate fondamentali volte a trovare modelli matematici in grado di riconoscere l'appartenenza a una classe degli oggetti sulla base di un insieme di misurazioni. La regola di classificazione k-Nearest Neighbor (k-NN) è concettualmente piuttosto semplice: un oggetto è classificato in base alle appartenenze di classe dei k oggetti più vicini, cioè è classificato in base alla maggioranza dei suoi k vicini più prossimi nello spazio dei dati. Pertanto, un oggetto è classificato sulla base della sua similarità con altri oggetti. Da un punto di vista computazionale, tutto ciò che è necessario è calcolare e analizzare una matrice di distanza. La distanza di ciascun oggetto da tutti gli altri oggetti viene calcolata e gli oggetti vengono quindi ordinati in base a questa distanza.Al fine di valutare quantitativamente gli effetti delle diverse geometrie indotte sui dati da ciascuna funzione di distanza, è stata eseguita l'analisi k-NN su 8 set di dati (Tabella 12). Non sono state adottate procedure di validazione sofisticate, ma solo la tipica tecnica implicita leave-one out dell'approccio k-NN. Pertanto, è stata stimata una valutazione approssimativa del potenziale comportamento di classificazione di ciascuna distanza per ciascun set di dati.

Il modo abituale di selezionare k è testare un insieme di valori k (ad esempio da 1 a 10); quindi, il k che dà il minor errore di classificazione può essere selezionato come quello ottimale.

Il Tasso di Non Errore (NER) è stato calcolato per ciascun set di dati mediante la media di ciascuna delle 18 misure di distanza insieme alla media del NER (Tabella 13). Il rango di ciascuna distanza per ogni set di dati è stato calcolato insieme al rango medio AR (Tabella 14).

Tabella 13. Tasso di Non Errore delle 18 distanze per ciascun set di dati e il Tasso di Non Errore medio (NER) dalla classificazione k-NN. In grigio, i migliori risultati per ciascun set di dati sono evidenziati.

Distanza	*Simbolo*	*Iris*	*Vini*	*Perpot*	*Sulfa*	*Thioph.*	*Itaoils*	*Sangue*	*Diabete*	$\overline{NER}$
Manhattan	UOMO	0.953	0.981	0.990	0.823	0.833	0.949	0.637	0.692	0.857
Euclidea	EUC	0.967	0.977	0.990	0.774	0.833	0.947	0.625	0.707	0.852

Soergel	SOE	0.953	0.967	0.980	0.788	0.792	0.947	0.637	0.708	0.847
Lance-Williams	LW	0.953	0.967	0.980	0.788	0.792	0.947	0.637	0.707	0.846
JT contrattato	CJT	0.967	0.981	0.990	0.760	0.792	0.946	0.626	0.708	0.846
Jaccard-Tanimoto	JT	0.960	0.972	0.980	0.724	0.792	0.947	0.628	0.718	0.840
Lagrange	LAG	0.967	0.955	0.970	0.774	0.792	0.943	0.609	0.697	0.838
Edge dell'Onda	NOI	0.953	0.972	0.960	0.788	0.750	0.929	0.633	0.663	0.831
Bhattacharyya	BHA	0.953	0.977	1.000	0.683	0.792	0.932	0.637	0.673	0.831
Canberra	PUÒ	0.947	0.977	0.960	0.752	0.750	0.929	0.637	0.672	0.828
Mahalanobis	MAH	0.913	0.972	0.990	0.710	0.792	0.920	0.626	0.700	0.828
Clark	CLA	0.953	0.986	0.970	0.732	0.750	0.911	0.632	0.662	0.825
LCM-simm. geom.	MSG	0.920	0.917	0.980	0.718	0.792	0.913	0.635	0.684	0.820
LCM-simm. aritm.	MSA	0.920	0.897	0.980	0.690	0.792	0.909	0.637	0.681	0.813
LCM-mat. inferiore.	ML	0.933	0.888	0.940	0.766	0.708	0.869	0.633	0.645	0.798
Coseno	CD	0.827	0.972	0.900	0.540	0.792	0.954	0.603	0.650	0.780
LCM-mater. superiore	MU	0.853	0.618	0.990	0.516	0.750	0.865	0.613	0.629	0.729
Correlazione	COR	0.853	0.980	0.500	0.518	0.708	0.947	0.598	0.629	0.717

Tabella 14. Ranks delle 18 distanze per ciascun set di dati e rango medio (AR) dalla classificazione k-NN. In grigio, il miglior rango(i) per ciascun set di dati è evidenziato.

Distanza	*Simbolo*	*Iris*	*Vini*	*Perpot*	*Sulfa*	*Tiop.*	*Itaoils*	*Diabete*	*del sangue*	*AR*
Manhattan	UOMO	7.5	2.5	4.0	1.0	1.5	2.0	3.5	8.0	3.8
Euclidea	EUC	2.0	6.0	4.0	5.5	1.5	5.0	14.0	3.5	5.2
JT contrattato	CJT	2.0	2.5	4.0	8.0	7.5	8.0	12.5	2.5	5.9
Soergel	SOE	7.5	12.5	9.0	3.0	7.5	5.0	3.5	2.5	6.3
Lance-Williams	LW	7.5	12.5	9.0	3.0	7.5	5.0	3.5	3.5	6.4
Jaccard-Tanimoto	JT	4.0	9.5	9.0	11.0	7.5	5.0	11.0	1.0	7.3
Bhattacharyya	BHA	7.5	6.0	1.0	15.0	7.5	10.0	3.5	11.0	7.7
Lagrange	LAG	2.0	14.0	12.5	5.5	7.5	9.0	16.0	7.0	9.2
Mahalanobis	MAH	15.0	9.5	4.0	12.0	7.5	13.0	12.5	6.0	9.9
Canberra	PUÒ	11.0	6.0	14.5	9.0	13.5	11.5	3.5	12.0	10.1
Edge dell'Onda	NOI	7.5	9.5	14.5	3.0	13.5	11.5	8.5	13.0	10.1
Clark	CLA	7.5	1.0	12.5	10.0	13.5	15.0	10.0	14.0	10.4
geometria simmetrica LCM	MSG	13.5	15.0	9.0	12.0	7.5	14.0	7.0	9.0	10.9
aritmetica simmetrica LCM	MSA	13.5	16.0	9.0	14.0	7.5	16.0	3.5	10.0	11.2
Coseno	CD	18.0	9.5	17.0	16.0	7.5	1.0	17.0	15.0	12.6
LCM-mat. inferiore.	ML	12.0	17.0	16.0	7.0	17.5	17.0	8.5	16.0	13.9
Correlazione	COR	16.5	4.0	18.0	17.0	17.5	5.0	18.0	17.5	14.2
LCM-mater. superiore	MU	16.5	18.0	4.0	18.0	13.5	18.0	15.0	17.5	15.1

Analizzando i risultati riportati nelle Tabelle 13 e 14, si può facilmente osservare che l'uso di diverse misure di distanza ha un effetto sulle prestazioni di classificazione, a causa del modo in cui ciascuna distanza rappresenta le relazioni di somiglianza/diversità tra gli oggetti. Cinque distanze forniscono i migliori risultati sia considerando il tasso medio di errore non (NER) che il rango medio (AR); queste sono le distanze Manhattan (MAN), Euclidea (EUC), Soergel (SOE), Jaccard-Tanimoto contratta (CJT) e Lance-Williams (LW).

Guardando i ranghi raccolti nella Tabella 14, solo in pochi casi il miglior rango (1. 0) è raggiunto da altre misure di distanza, come la distanza di Clark (CLA) per il set di dati Wines, Bhattacharyya (BHA) per il set di dati Perpot, la distanza coseno (CD) per il set di dati Itaoils e Jaccard-Tanimoto (JT) per il set di dati Diabetes.

La distanza Jaccard-Tanimoto contratta (CJT) è migliore della sua distanza Jaccard-Tanimoto parentale (JT) in 4 set di dati su 8, in 1 caso è uguale e in 3 casi è solo leggermente inferiore.

Le quattro distanze di tipo Mahalanobis derivate dalla distanza LCM (MSA, MSG, MU, ML) sembrano non essere utili nei problemi di classificazione; d'altra parte, la distanza LCM è stata proposta per scopi diversi, cioè con l'obiettivo di individuare outlier e analizzare il dominio di applicabilità di un modello.

La correlazione e la distanza coseno sembrano essere molto deboli per scopi di classificazione nei set di dati studiati; considerazioni analoghe possono essere fatte per la distanza di Canberra (CAN), Wave-Edge (WE) e la distanza di Clark (CLA), sebbene quest'ultima abbia fornito il miglior risultato in un caso.

Un esempio dettagliato di come la scelta della funzione di distanza influisca sulla geometria dello spazio degli oggetti è fornito per il set di dati Vini (Figura 21). Il confronto è stato effettuato tramite Scaling Multidimensionale. Come commentato in precedenza, la distanza di Clark (CLA) fornisce i migliori risultati (NER = 0.986) e questo è visivamente confermato dalla buona clusterizzazione degli oggetti appartenenti alle tre diverse classi, che non è così ben ottenuta nel caso della distanza Euclidea (EUC, NER = 0. 977), Jaccard-Tanimoto (JT, NER = 0.972) e Jaccard-Tanimoto contratta (CJT, NER = 0.981).

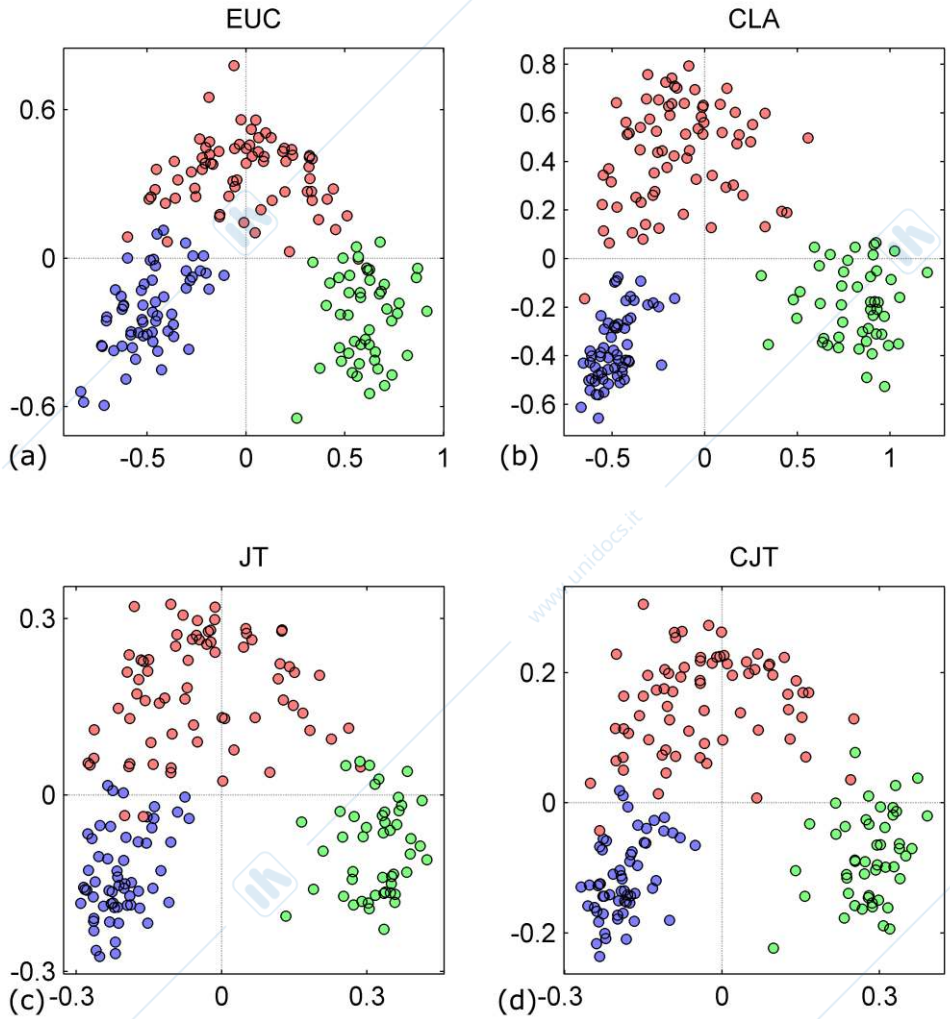

Figura 21 Proiezioni dei 178 oggetti di vino mediante Scaling Multi-Dimensionale basato sulle distanze Euclidee (EUC, Figura 15a), Clark (CLA, Figura 15b), Jaccard-Tanimoto (JT, Figura 15c) e Jaccard-Tanimoto Contratto (CJT, Figura 15d). I diversi colori rappresentano le classi.

8 Confronto dei coefficienti di similarità binaria

È stato effettuato un confronto esteso tra i 44 misure di similarità elencate nella Tabella 6 utilizzando un set di dati simulato [13]. Il set di dati simulato ha generato 100.000 casi creando casualmente quadruple di numeri interi (a, b, c, d) sotto il vincolo a + b + c + d = 1024. Per ogni caso, sono stati calcolati i 44 coefficienti di similarità e organizzati in una matrice di 100.000 righe e 44 colonne. Ogni caso può essere considerato come il confronto di un vettore binario di lunghezza 1024 bit con un vettore di riferimento della stessa lunghezza.

I 100.000 valori di similarità generati per ciascun coefficiente sono stati analizzati per calcolare le seguenti statistiche descrittive: media, deviazione standard (std), coefficiente di variazione (cv), percentili 5 e 95 (perc(5) e perc(95)). Questi valori sono elencati nella Tabella 15. I valori minimi e massimi di tutti i coefficienti sono rispettivamente 0 e 1.

L'ispezione della Tabella 15 suggerisce che la maggior parte dei coefficienti ha un valore medio attorno a 0,5 e che coprono in modo soddisfacente l'intervallo di similarità. Ci sono tre coefficienti molto anomali: Co1 (B24), Co2 (B25) producono valori molto elevati, e Mou (B16) produce valori molto bassi. Questi coefficienti anomali sono probabilmente stati originariamente proposti per trattare vettori brevi, dove i parametri b, c, d possono avere meno influenza rispetto a a. Un comportamento meno estremo è mostrato da CT1 (B39), CT4 (B42), CT3 (B41), Sim (B7), SS2 (B13), Di2 (B32) e Kul (B11) (che hanno tutti valori medi superiori a 0,55) e CT2 (B40), SS4 (B29), HL (B38) e GK (B27) (che hanno tutti valori medi inferiori a 0,30).Passando alle deviazioni standard (escludendo Co1 (B24), Co2 (B25) e Mou (B16)), i coefficienti che mostrano la massima variabilità sono Yu1 (B20), Sor (B33), Di2 (B32), Di1 (B31), Fos (B22), For (B6), Gle (B4), Ja (B14), BB (B8), JT (B3) e DK (B9) (tutti con deviazioni standard superiori a 0,30), mentre la minima variabilità è fornita da Den (B23), CT1 (B39), CT5 (B43), CT2 (B40), dis (B26) e Coh (B34) (tutti con deviazioni standard inferiori a 0,20).

Tabella 15. Parametri statistici per 44 coefficienti di similarità calcolati dal set di dati simulato.

#	Simbolo	media	___	cv	perc(5)	perc(95)
B1	SM	0.5420	0.2869	0.5293	0.0625	0.9629
B2	RT	0.4268	0.2848	0.6672	0.0323	0.9284
B3	JT	0.4050	0.3112	0.7685	0.0102	0.9433
B4	Gle	0.5066	0.3204	0.6325	0.0201	0.9708
B5	RR	0.3329	0.2939	0.8829	0.0059	0.9014
B6	Per	0.3756	0.3234	0.8609	0.0015	0.9429
B7	Sì	0.6902	0.2984	0.4323	0.0865	0.9971
B8	BB	0.4427	0.3185	0.7194	0.0119	0.9569

B9	DK	0.5302	0.3075	0.5800	0.0391	0.9710
B10	BUB	0.5052	0.2980	0.5898	0.0375	0.9569
B11	Kul	0.5665	0.2875	0.5075	0.0606	0.9712
B12	SS1	0.3092	0.2880	0.9312	0.0051	0.8927
B13	SS2	0.6532	0.2716	0.4158	0.1176	0.9811
B14	Sì	0.5658	0.3187	0.5633	0.0299	0.9804
B15	Fai	0.4374	0.2655	0.6069	0.0454	0.9170
B16	Mou	0.0085	0.0502	5.9205	0.0001	0.0232
B17	Mic	0.5099	0.2155	0.4227	0.1096	0.8912
B18	RG	0.4558	0.2411	0.5290	0.0611	0.8669
B19	HD	0.3619	0.2353	0.6503	0.0328	0.8031
B20	Yu1	0.5349	0.3700	0.6917	0.0025	0.9978
B21	Yu2	0.5252	0.2837	0.5401	0.0479	0.9551
B22	Fos	0.3746	0.3234	0.8635	0.0013	0.9425
B23	Den	0.3431	0.1336	0.3894	0.1270	0.5880
B24	Co1	0.9921	0.0803	0.0809	0.9970	0.9999
B25	Co2	0.9921	0.0802	0.0809	0.9966	0.9999
B26	dis	0.5066	0.1563	0.3085	0.2244	0.7766
B27	GK	0.2961	0.2579	0.8710	0.0049	0.8076
B28	SS3	0.5213	0.2161	0.4146	0.1475	0.8748
B29	SS4	0.2465	0.2410	0.9778	0.0018	0.7527
B30	Phi	0.5175	0.2140	0.4136	0.1379	0.8727
B31	Di1	0.5473	0.3248	0.5934	0.0256	0.9868
B32	Di2	0.5856	0.3390	0.5788	0.0229	0.9946
B33	Sor	0.4051	0.3449	0.8516	0.0007	0.9737
B34	Coh	0.5378	0.1788	0.3325	0.2461	0.8673
B35	Pe1	0.5210	0.2308	0.4429	0.1157	0.9014
B36	Pe2	0.5167	0.2191	0.4240	0.1214	0.8899
B37	MP	0.5164	0.2105	0.4076	0.1397	0.8708
B38	HL	0.2687	0.2102	0.7824	0.0270	0.7124

B39	CT1	0.8733	0.1354	0.1551	0.6022	0.9946
B40	CT2	0.1628	0.1525	0.9364	0.0093	0.4715
B41	CT3	0.7401	0.2208	0.2983	0.2807	0.9850
B42	CT4	0.7734	0.2160	0.2793	0.3197	0.9914
B43	CT5	0.5128	0.1400	0.2729	0.2857	0.7384
B44	CA	0.5316	0.2174	0.4090	0.1609	0.8766

Le sequenze ordinate dei valori di similarità (in ordine crescente) sono state tracciate per ciascun coefficiente per esplorare la forma funzionale. Per semplificare l'analisi e la discussione che seguono, i grafici sono presentati in tre figure diverse: funzioni simmetriche (Figura 22), funzioni asimmetriche (Figura 23) e funzioni basate sulla correlazione (Figura 24). L'ispezione di queste figure mostra che le forme delle funzioni possono essere approssimativamente categorizzate come logaritmiche, esponenziali, sigmoidi o quasi lineari nel carattere.

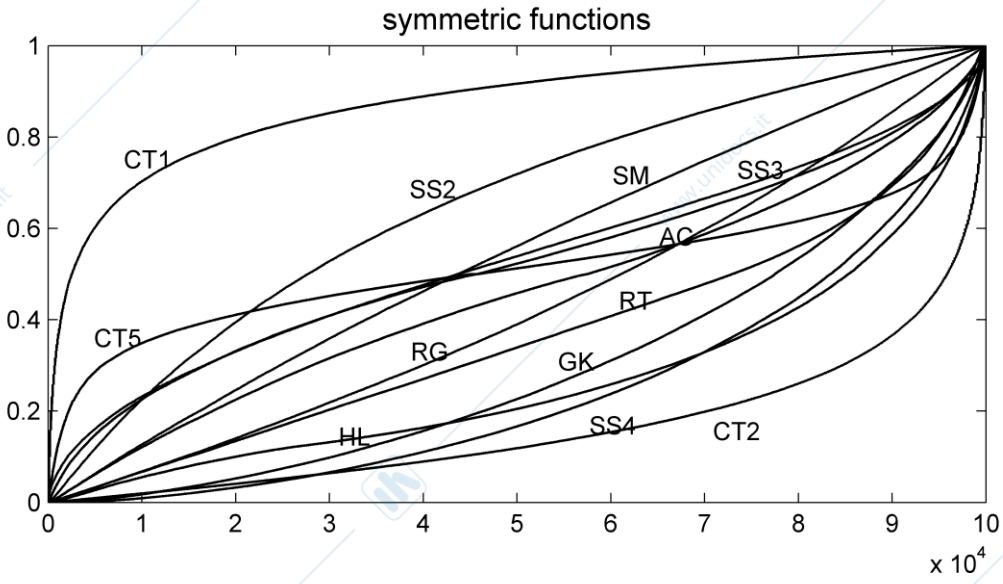

Figura 22 Grafici a linee dei coefficienti binari simmetrici calcolati dal dataset simulato.

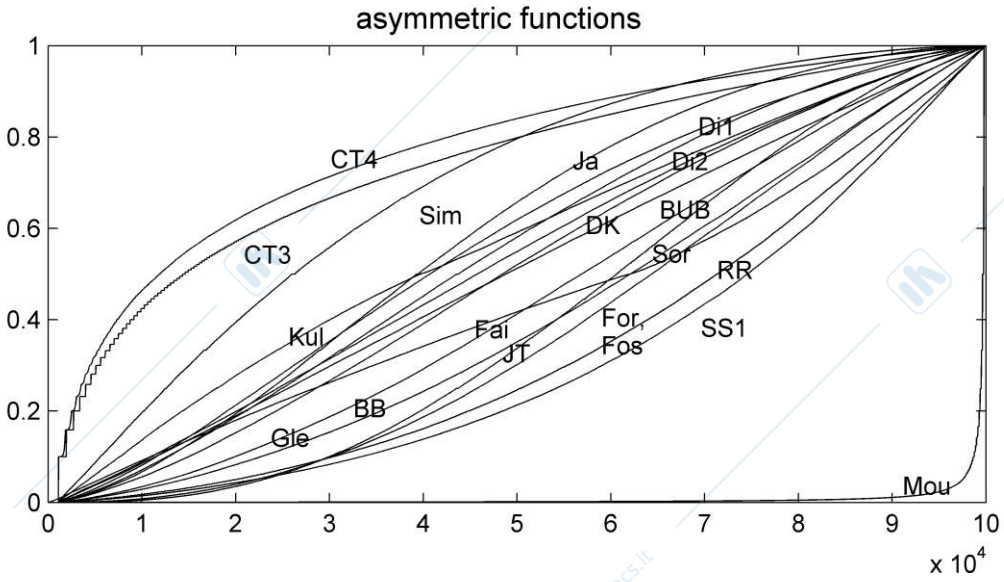

Figura 23 Grafici a linee dei coefficienti binari asimmetrici calcolati dal dataset simulato.

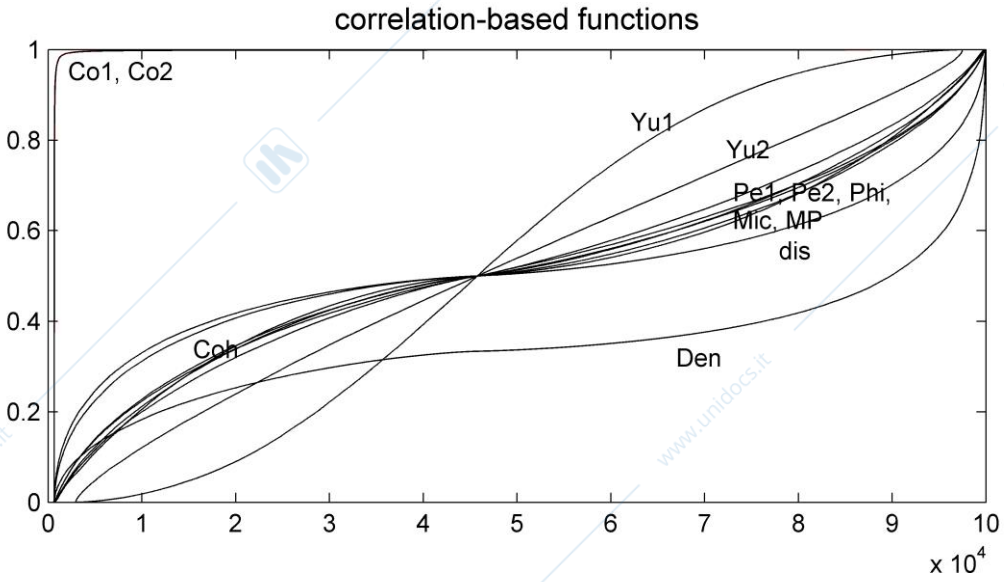

Figura 24 Grafici a linee dei coefficienti binari basati sulla correlazione calcolati dal dataset simulato.

La Scalatura Multi-Dimensionale è stata eseguita sulla matrice 41x41 dei coefficienti di correlazione di Pearson calcolati dai dati simulati. Questa analisi omette i coefficienti di Mountford e Cole (cioè, Co1 (B24), Co2 (B25) e Mou (B16)) poiché sono outlier significativi (vedi Figure 23 e 24). La configurazione finale dei coefficienti di similarità binaria in un grafico MDS bidimensionale è mostrata nella Figura 25.

A prima vista, i coefficienti di similarità sembrano ben raggruppati secondo le loro proprietà di simmetria, con le funzioni simmetriche (quadrati verdi, sul lato medio-sinistro in basso), le funzioni asimmetriche (triangoli blu, sul lato destro) e le funzioni basate sulla correlazione (cerchi rossi, sul lato sinistro in alto) ben separate l'una dall'altra. A questo proposito, è interessante notare che BUB (B10) e Fai (B15), che sono intermedi tra funzioni simmetriche e asimmetriche, sono opportunamente collocati tra i cluster simmetrici e asimmetrici.

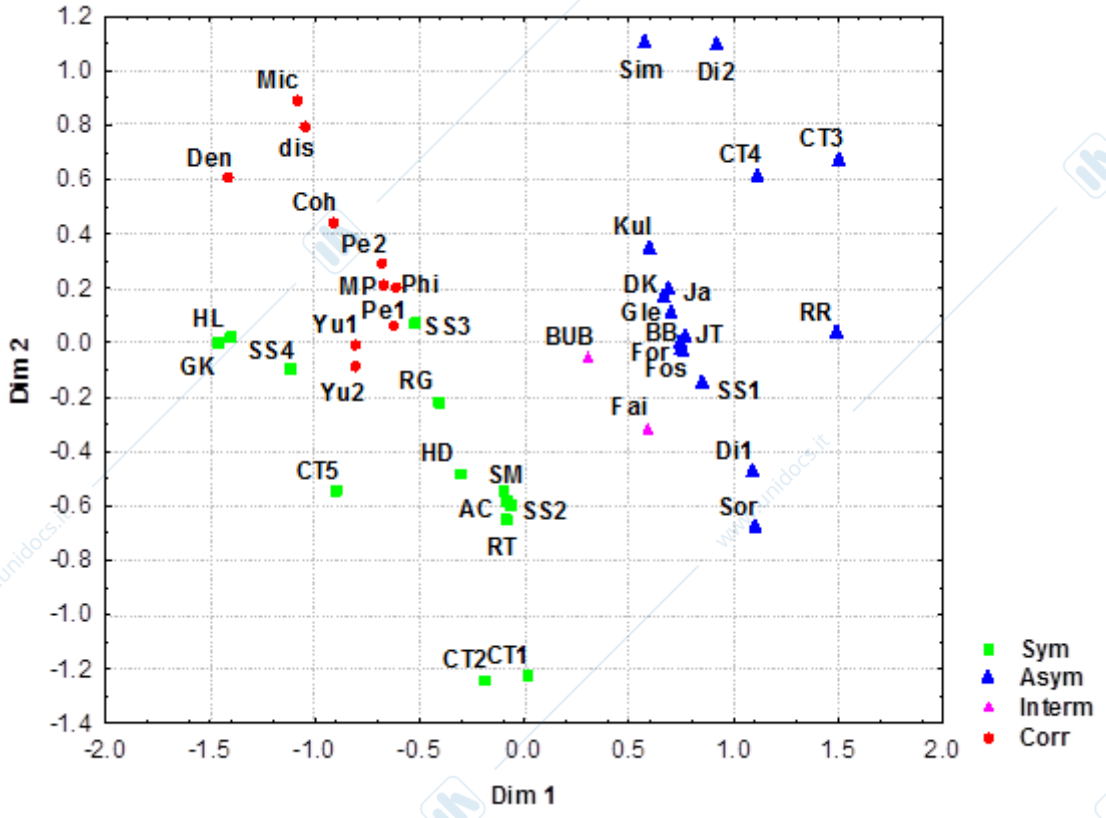

Figura 25 Scalatura multidimensionale dei coefficienti di similarità binaria. I coefficienti Co1, Co2 e Mou sono stati esclusi dall'analisi poiché sono forti outlier (vedi testo).

Molti dei coefficienti sono molto vicini tra loro nel grafico, indicando relazioni di similarità ravvicinate, ad esempio, il gruppo composto da SM (B1), RT (B2), SS2 (B13) e AC (B44), che hanno una correlazione di rango pari a uno. Allo stesso modo, il gruppo JT (B3), Ja (B14), Gle (B4), SS1

(B12), Per (B6), Fos (B22) e DK (B9) hanno correlazioni di rango superiori a 0, 99. Alcuni coefficienti, tuttavia, sono piuttosto isolati nel grafico MDS. Questo è il caso delle coppie CT1 (B39) e CT2 (B40), Sim (B7) e Di2 (B32), CT3 (B41) e CT4 (B42); RR (B5) e CT5 (B43) sembrano anche essere piuttosto separati dagli altri coefficienti.

8 Riconoscimenti

Per la prima edizione di questo lavoro, gli Autori ringraziano calorosamente Michel Marie Deza dell'Ecole Normal Supérieure (Parigi, Francia) per il suo aiuto nella gestione delle proprietà matematiche delle distanze e Luis Peinador Sarabia dell'Università di Burgos (Spagna) per i suoi suggerimenti.

Questa seconda edizione è dedicata alla memoria di Michel Marie Deza (1939 – 2016).

9 Riferimenti

1. W. J. Krzanowski, Principi dell'Analisi Multivariata, pp. 564, 1988.
2. R. Todeschini, D. Ballabio, M. Cassotti e V. Consonni, N3 e BNN: Due nuovi metodi di classificazione basati sulla similarità in confronto con altri classificatori, J.Chem.Inf.Model., 55, 2365-2375 (2015).
3. F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni e R. Todeschini, Confronto di diversi approcci per definire il dominio di applicabilità dei modelli QSAR, Molecole, 17, 4791-4810 (2012).
4. V. Batagelj e M. Bren, Confronto delle misure di somiglianza, J.Classif., 12, 73-90 (1995).
5. M. M. Deza e E. Deza, Enciclopedia delle Distanze, pp. 590, 2009.
6. P. Legendre e L. Legendre, Ecologia Numerica (2a Ed.), pp. 854, 1998.
7. P. H. A. Sneath e R. R. Sokal, Tassonomia Numerica, 1973.
8. S.-H. Cha, Indagine Completa sulle Misure di Distanza/Somiglianza tra funzioni di densità di probabilità, Giornale Internazionale di Modelli Matematici e Metodi nelle Scienze Applicate, 1, 300-307 (2007).
9. C. M. Cuadras, Distanze Statistiche (in spagnolo), Statistica Spagnola, 30, 295-378 (1989).
10. A. G. Maldonado, J. P. Doucet, M. Petitjean e B. T. Fan, Somiglianza e diversità molecolare in chemoinformatica: dalla teoria alle applicazioni, Mol.Div., 10, 39-79 (2006).

11. Y. C. Martin, J. L. Kofron e L. M. Traphagen, Le molecole strutturalmente simili hanno attività biologica simile?, J.Med.Chem., 45, 4350-4358 (2002).

12. P. Willett, Somiglianza e clustering nei sistemi di informazione chimica, pp. 254 (1987).

13. R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema e P. Willett, Coefficienti di somiglianza per dati chemoinformatici binari: panoramica e confronto esteso utilizzando dataset simulati e reali, J.Chem.Inf.Model., 52, 2884-2901 (2012).

14. T. Fearn, C. Riccioli, A. Garrido-Varo e J. E. Guerrero-Ginel, Sulla geometria di SNV e MSC, *Chemometria & Sistemi di Laboratorio Intelligenti, 96, 22-26 (2009).*

15. J. Engel, J. Gerretzen, E. Szymanska, J. J. Jansen, G. Downey, L. Blanchet e L. M. C. Buydens, Rompere con le tendenze nella pre-elaborazione?, Tendenze nella Chimica Analitica, 50, 96-106 (2013).

16. P. Filzmoser, K. Hron e C. Reimann, Analisi delle componenti principali per dati composizionali con outlier, Environmetrics, 20, 621-632 (2009).

17. D. Weinshall, D. W. Jacobs e Y. Gdalyahu, Classificazione in spazi non metrici, p. 837-844 (1998).

18. R. Todeschini, D. Ballabio, V. Consonni, F. Sahigara e P. Filzmoser, Distanza Mahalanobis centrata localmente: una nuova misura di distanza con caratteristiche salienti per la rilevazione di outlier, Anal.Chim.Acta, **787**, 1-9 (2013).

19. M. Dehmer, F. Emmert-Streib e Y. Shi, Interrelazioni delle misure di distanza nei grafi basate su indici topologici, PLOS One, 9, e94985 (2014).

20. U. Fechner e G. Schneider, Valutazione delle metriche di distanza per la ricerca di somiglianze basata sui ligandi, *ChemBioChem*, **5**, 538-540 (2004).

21. F. Decheng, S. Jon, C. Pang, W. Dong e C. Won, Analisi di clustering quantistico migliorata basata sulla distanza ponderata e la sua applicazione, Helyon, 4, e00984-(2018).

22. P. Diaconis e R. L. Graham, La regola del piede di Spearman come misura di disordine, Journal of the Royal Statistical Society (Serie B), 39, 262-268 (1977).

23. K. HebÈrger, La somma delle differenze di ranking confronta metodi o modelli in modo equo, Trends in Analytical Chemistry, 29, 101-109 (2010).

24. J. A. Koziol, Somme delle differenze di ranking e numeri di inversione per la discriminazione dei metodi, *J.Chemom.*, **27**, 165-169 (2013).

25. D. Bajusz, A. Racz e K. HebÈrger, Perché l'indice di Tanimoto è una scelta appropriata per i calcoli di similarità basati su impronte digitali?, J.Cheminf., 7, 1-13 (2015).

26. J. H. Zar, Analisi Biostatistica, pp. 718 (1984).

27. D. E. Rumelhart, P. Smolensky, J. L. McClelland e G. E. Hinton, Schemata e processi di pensiero sequenziali nei modelli PDP, 1986.

28. F. Grisoni, V. Consonni e D. Ballabio, Consenso di Machine Learning per prevedere il legame con il recettore androgeno all'interno del progetto CoMPARA., J.Chem.Inf.Model., 59, 1839-1848 (2019).

29. R. Todeschini, D. Ballabio, F. Grisoni e V. Consonni, Un nuovo concetto di similarità di secondo ordine e il ruolo delle misure di distanza/similarità nei metodi di classificazione locale, Chemometrics & *Intell.Lab.Syst.*, **157**, 50-57 (2016).

30. A. Mauri, D. Ballabio, R. Todeschini e V. Consonni, Miscele, metaboliti, liquidi ionici: una nuova misura per valutare la similarità tra sistemi chimici complessi, J.Cheminf., 8, 1-3 (2016).

31. J. C. Gower, Analisi Procruste generalizzata, Psychometrika, 40, 31-51 (1975).

32. R. Todeschini, V. Consonni, A. Manganaro, D. Ballabio e A. Mauri, Misura Canonica di Correlazione (CMC) e Misura Canonica di Distanza (CMD) tra insiemi di dati. Parte 1. Teoria e semplici applicazioni chemometriche, Anal.Chim.Acta, 648, 45-51 (2009).

33. R. Todeschini, Correlazione dei dati, numero di componenti principali significativi e forma delle molecole. L'indice di correlazione K., Anal.Chim.Acta, 348, 419-430 (1997).

34. R. Todeschini, V. Consonni e A. Maiocchi, L'indice di correlazione K: sviluppo della teoria e le sue applicazioni in chemometria., Chemometrics & Intell.Lab.Syst., 46, 13-29 (1999).

35. R. A. Fisher, L'uso di misurazioni multiple nei problemi tassonomici, Ann.Eugenics, 7, 179-188 (1936).

36. M. Forina, C. Armanino, S. Lanteri ed E. Tiscornia, Classificazione degli oli d'oliva in base alla loro composizione di acidi grassi (1983).

37. M. Forina, Dataset artificiale di M. Forina (Univ. di Genova), 2000.

38. Y. Miyashita, Y. Takahashi, C. Takayama, T. Ohkubo, K. Fumatsu e S. Sasaki, Studi assistiti da computer sulla struttura/gusto dei sulfamati mediante metodi di riconoscimento dei modelli, Anal. Chim. Acta, 184, 143-149 (1986).

39. P. P. Mager, Statistiche di design in farmacochimica, p. 493-493 (1991).

40. M. Forina, C. Armanino, M. Castino e M. Ubigli, Analisi dei dati multivariati come metodo discriminante dell'origine dei vini, Vitis, 25, 189-201 (1986).

41. K. A. Baggerly, J. S. Morris, S. R. Edmonson e K. R. Coombes, Segnale nel rumore: Valutazione della riproducibilità riportata dei test proteomici sierici per il cancro ovarico, Journal of National Cancer Institute, **97**, 307-309 (2005).

42. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler e R. S. Johannes, Utilizzo dell'algoritmo di apprendimento ADAP per prevedere l'insorgenza del diabete mellito, Atti del Simposio su Applicazioni Informatiche e Cure Mediche, 9, 261-265 (1988).