

Pragmatica 2001-02-19

In questo lavoro vengono analizzati varie tipologie di distanze e pseudodistanze definite in un universo. Lo scopo è quello di definire metodiche utili al trattamento di insiemi di informazioni in cui sia presente un meccanismo di memorizzazione e ricerca che tenga conto della "semantica" delle informazioni date.

Per questo fine si intende definire una distanza (o una pseudodistanza) tra le informazioni; infatti, la memorizzazione e la ricerca delle informazioni richiede che si abbiano dei metodi per individuare una "somiglianza" tra due informazioni; questo può essere appunto fatto utilizzando la distanza per creare degli intorni di un'informazione data; di fatto si può scegliere un intorno più o meno grande, creando quindi una gerarchia delle somiglianze.

La distanza tra informazioni può essere definita in maniera relativamente semplice. È peraltro utile poter operare su informazioni complesse dotate di un'organizzazione intermedia. In questo caso un metodo preliminare di rappresentazione delle informazioni è basato sul considerarle come insiemi di informazioni elementari, che a loro volta possiedono una metrica. Nel modello preliminare qui definito, queste informazioni sono rappresentate da insiemi definiti in un universo di insiemi di insiemi. Come si vedrà, in questo modello si misurerà la distanza tra informazioni non attraverso una distanza ma per mezzo di una pseudodistanza.

Pseudodistanza 2001-02-19

Ricordiamo la definizione di una distanza. Si dice che d è una pseudodistanza in U se per ogni $A, B \in U$ si ha:

1. $d(A, B) = 0$ se e solo se $A = B$ (proprietà di unicità delle distanze);
2. $d(A, B) = d(B, A)$ (proprietà riflessiva);
3. $d(A, B) \leq d(A, C) + d(C, B)$ (proprietà triangolare).

È importante notare che, in conseguenza degli assiomi dati, per ogni $A, B \in U$ si ha $d(A, B) \geq 0$. Si noti che l'assioma 1 può essere indebolito nella forma $d(A, A) = 0$. In questo caso si può ottenere una struttura che non è una vera e propria distanza ma che può essere trasformata in una distanza considerando le classi di equivalenza $X^v = \{Y \mid d(Y, X) = 0\}$.

Definiamo la struttura algebrica pseudodistanza. Diremo che D è una pseudodistanza in U se per ogni $A, B \in U$ si ha:

1. $D(A, B) = D(A, A)$ se e solo se $A = B$ (proprietà di unicità delle pseudodistanze);
2. $D(A, B) = D(B, A)$ (proprietà riflessiva);
3. $D(A, B) \leq D(A, C) + D(C, B)$ (proprietà triangolare).

Si può dimostrare che, per ogni per ogni $A, B \in U$ si ha $0 \leq D(A, A) \leq D(A, B)$. Ne segue inoltre facilmente che $D(A, B) \leq \min\{D(A, A), D(B, B)\}$.

Anche in questo caso la prima condizione può essere indebolita, nella forma $D(A, A) \leq D(A, B)$. Si noti che, dato questa condizione più debole e ipotizzando che sia una relazione di equivalenza $X \sim Y \iff D(X, Y) = 0$, si ottiene la condizione più forte. Non è però vero in generale che la relazione data sia di equivalenza. Un controesempio è il seguente: si abbiano tre elementi x, y, z , con la seguente tabella di pseudodistanze:

	x	y	z
x	2	2	3
y	2	2	2
z	3	2	2

Si vede subito che sono soddisfatti gli assiomi 2 e 3 e la condizione alternativa dell'assioma 1. Non è però definibile la classe di equivalenza in quanto si avrebbe che $x \sim y$ e $y \sim z$ ma $x \not\sim y$. Infatti $d(x, z) = 3 \cdot 2 = d(x, x) = d(z, z)$. È possibile però aggiungere un ulteriore assioma $d(A, B) \leq \frac{d(A, A) + d(B, B)}{2}$. In presenza di questo assioma e dell'assioma indebolito, la relazione $\sim$ precedentemente descritta è effettivamente una relazione di equivalenza. Si noti che nel caso delle distanza, questo assioma aggiuntivo non è necessario in quanto è una conseguenza degli altri. Per dimostrare che si tratta effettivamente di una relazione di equivalenza, essendo ovvio che $X \sim X$ e che, se $X \sim Y$ allora $Y \sim X$, basta dimostrare la proprietà transitiva. Sia $X \sim Y$ e $Y \sim Z$ e si voglia dimostrare che $X \sim Z$. È immediato che $d(X, X) = d(Z, Z)$. D'altra parte si ha che $d(X, Z) \leq \frac{d(X, X) + d(Z, Z)}{2} = d(X, X)$ e $d(X, Z) \leq \min\{d(X, X), d(Z, Z)\}$, da cui la tesi.

Si dimostra immediatamente che in ogni U vi è un $m_U \geq 0$ tale che $m_U \leq d(A, A)$ per ogni $A \in U$; basta scegliere $m_U = \min_{A \in U} d(A, A)$. Non è peraltro detto che vi sia un $A \in U$ tale che $m_U = d(A, A)$.

Distanza su insiemi 2001-02-19

Sia disponibile un universo U di simboli, atti a descrivere informazioni. Nell'universo U possono essere definite varie distanze. Uno dei metodi consiste nel definire una distanza sulla base di una metrica sulle informazioni, ovvero utilizzando una funzione di misura W . Avendo una misura W si può infatti definire una distanza d nel modo seguente:

$$d(A, B) = k \cdot \frac{W(A \# B)}{W(A) W(B)} = k \cdot \left(1 - \frac{W(A \cap B)}{W(A) W(B)} \right)$$

dove $\#$ è la differenza simmetrica di insiemi.

Una possibile (ma non l'unica) definizione della misura W è quella che segue; sia data una funzione reale positiva W^0 qualunque definita sugli elementi di U ; data $W^0 : U \rightarrow \mathbb{R}^+ =]0, \infty[$ si definirà

$$W(A) = \sum_{a \in A} W^0(a)$$

Si noti che $d(A, A) = k$ per ogni $A \in U$. Va invece definita a parte la distanza $d(\emptyset, \emptyset) = 0$.

La definizione di d può essere estesa al caso infinito (numerabile) come segue; sia $A = \{a_1, a_2, \dots, a_m, \dots\}$ e $B = \{b_1, b_2, \dots, b_m, \dots\}$; posto $A_m = \{a_1, a_2, \dots, a_m\}$ e $B_m = \{b_1, b_2, \dots, b_m\}$, si definirà

$$d(A, B) = \lim_{m \rightarrow \infty} d(A_m, B_m)$$

Si noti che, in base alla definizione data, $0 \leq d(A_m, B_m) \leq k$ per ogni m , e quindi anche $0 \leq d(A, B) \leq k$.

Per l'estensione al caso infinito più che numerabile, possono essere considerate altre W . Una possibile estensione è quella in cui $W^0 : U \rightarrow]0, \infty[$; in questo caso esistono coppie di insiemi diversi la cui distanza è nulla, in quanto gli elementi x tali che $W^0(x) = 0$ non portano nessuna informazione. Si avrebbe infatti che $d(A, A \cup \{x\}) = k \cdot \left(1 - \frac{W(A \cap \{x\})}{W(A) W(A \cup \{x\})} \right)$ dove è $W(A \cap \{x\}) = W(A) + W(x) = W(A)$ e $W(A \cup \{x\}) = W(A) + W(x)$ per cui $d(A, A \cup \{x\}) = 0$.

Insiemi di insiemi 2001-04-09

Il trattamento degli insiemi definiti al punto precedente è abbastanza soddisfacente volendo trattare unicamente informazioni elementari non strutturate. Quando si voglia però passare a informazioni strutturate, si hanno vari casi in cui una definizione del tipo precedentemente

descritta non sembra ottimale. A questo scopo si vuole analizzare la possibilità di operare su semplici informazioni dotate di struttura. In questo caso si farà riferimento a una struttura basata sul concetto di insieme. Le informazioni saranno cioè considerate come insiemi di insiemi.

In quest'ottica, infatti, si distinguerà tra **simboli elementari** (per esempio i caratteri) che non sono scomponibili in nessun modo e **simboli complessi** (per esempio le parole, le frasi, le immagini) che sono qui considerati insiemi di simboli elementari. L'obiettivo è quello di operare su questi simboli complessi. L'ipotesi che i simboli complessi siano formati da simboli elementari è in realtà un'ipotesi riduttiva, in quanto un'immagine è formata da altre immagini (più semplici) in un processo che di fatto potrebbe non arrivare a termine.

Definiremo quindi un insieme E di simboli elementari $a, b, c, \dots$ e un insieme C di simboli complessi $X, Y, Z, \dots$ tali che per ogni $X \in C$ sia $X \subseteq E$. Quelli che interessano sono gli insiemi X di simboli complessi tali che $X \subseteq C$. Si utilizzerà, quando comodo, la notazione

$$X = \{X_1, \dots, X_m\} \quad \hat{X} = \{X_i \mid i = 1, \dots, m\}$$

o, quando non ambiguo, $X = \hat{X}$. Chiameremo simboli atomici quei simboli complessi che contengono un unico elemento (ovvero, per i quali è $m = 1$).

Siano dati allora due insiemi di simboli complessi $A = \{A_1, \dots, A_m\}$ e $B = \{B_1, \dots, B_n\}$. Si definisce allora la funzione

$$D(A, B) = \frac{\sum_{i=1, \dots, m} \sum_{j=1, \dots, n} e_{A, B}(A_i, B_j)}{m \times n}.$$

dove

$$e_{A, B}(A_i, B_j) = \begin{cases} 0 & \text{se } A_i \notin A \text{ o } B_j \notin B \text{ e } A_i \notin B \text{ o } B_j \notin A \\ \frac{1}{2} & \text{se } \{A_i, B_j\} \subseteq A \cup B \\ 1 & \text{altrimenti} \end{cases}$$

Si noti che $e_{A, B}(A_i, B_j) = e_{A, B}(B_j, A_i) = e_{B, A}(A_i, B_j) = e_{B, A}(B_j, A_i)$. Si noti che in nessun modo questa definizione è influenzata dal fatto che gli insiemi $A_1, \dots, A_m, B_1, \dots, B_n$ siano di cardinalità finita o no. Infatti la cardinalità infinita può al più influenzare la distanza d . Purché quest'ultima sia correttamente definita, la definizione della funzione D è comunque ben posta. Più avanti accenneremo invece al caso in cui siano infinite le cardinalità di A e B .

Diamo alcuni esempi basati sulla distanza d e sulla misura W definite come nel precedente paragrafo.

¾ Siano dati i simboli atomici $A = \{a, b\}$, $B = \{a, x\}$ e $C = \{x, y\}$. Posto $A = \{a, b\}$, si avranno: $D(A, A) = \frac{\frac{1}{2}d(A, A)}{1} = \frac{W(A, A)}{W(A, A)} = \frac{0}{2} = 0$; $D(A, B) = \frac{2}{3}$; $D(A, C) = \frac{4}{4} = 1$.

¾ Siano $A = \{A_1, A_2\}$ e $B = \{B_1, B_2\}$, con $A_1 = \{a, b\}$, $A_2 = \{c, d\}$, $B_1 = \{a, c\}$, $B_2 = \{b, d\}$. Si può notare che, per ogni A_i, B_j è $e_{A, B}(A_i, B_j) = 1$. Si ottiene quindi

$$D(A, B) = \frac{\sum_{i,j} 1}{2 \times 2} = \frac{d(A_1, B_1) + d(A_1, B_2) + d(A_2, B_1) + d(A_2, B_2)}{4} = \frac{\frac{2}{3} + \frac{2}{3} + \frac{2}{3} + \frac{2}{3}}{4} = \frac{2}{3}.$$

¾ Sia $A = \{A_1, A_2\}$, con $A_1 = \{a, b\}$, $A_2 = \{c, d\}$. Si può notare che, per ogni A_i, A_j è $e_{A, A}(A_i, A_j) = \frac{1}{2}$. Si ottiene quindi

$$D(A, A) = \frac{\sum_{i,j} \frac{1}{2}}{2 \times 2} = \frac{\frac{1}{2}d(A_1, A_1) + \frac{1}{2}d(A_1, A_2) + \frac{1}{2}d(A_2, A_1) + \frac{1}{2}d(A_2, A_2)}{4} = \frac{0 + \frac{1}{2} + \frac{1}{2} + 0}{4} = \frac{1}{4}.$$

È interessante notare che $D(A, A) > 0$ se e solo se $|A| > 1$. Abbiamo già osservato che se $|A| = 1$ allora $D(A, A) = 0$. Basta allora considerare $A = \{A_1, A_2\}$, con $A_1 \cap A_2$. Si ha immediatamente che $D(A, A) \geq \frac{d(A_1, A_2)}{|A|^2} > 0$.

Si può dimostrare che la funzione D descritta è una pseudodistanza. Osserviamo innanzitutto che $\sum_{A \in \mathcal{A}, B \in \mathcal{B}, C \in \mathcal{C}} d(A, B) = |\mathcal{C}| \sum_{A \in \mathcal{A}, B \in \mathcal{B}} d(A, B)$.

Possiamo allora anticipare il seguente lemma:

Se $\mathcal{A}$ e $\mathcal{B}$ sono disgiunti, è $D(\mathcal{A}, \mathcal{A}) < D(\mathcal{A}, \mathcal{B})$.

• Se $|\mathcal{A}| = 1$ il lemma è ovvio, in quanto $D(\mathcal{A}, \mathcal{A}) = 0$ ed, essendo $\mathcal{A}$ e $\mathcal{B}$ disgiunti, $D(\mathcal{A}, \mathcal{B}) > 0$. Sia quindi $|\mathcal{A}| = n > 1$. Si avrà allora che

$$\begin{aligned} D(\mathcal{A}, \mathcal{B}) &= \frac{\sum_{i,j} d(A_i, B_j)}{|\mathcal{A}| |\mathcal{B}|} \\ &= \frac{\sum_{i,j,k} d(A_i, B_j) + d(B_j, A_k)}{|\mathcal{A}| |\mathcal{B}| |\mathcal{A}|} \\ &> \frac{\sum_{i,j,k} d(A_i, A_k)}{|\mathcal{A}| |\mathcal{B}| |\mathcal{A}|} \\ &= \frac{\sum_{i,k} d(A_i, A_k)}{|\mathcal{A}| |\mathcal{A}|} \\ &> \frac{1}{2} \frac{\sum_{i,k} d(A_i, A_k)}{|\mathcal{A}|} \\ &= \frac{1}{2} D(\mathcal{A}, \mathcal{A}) \end{aligned}$$

Dal fatto che $|\mathcal{A}| > 1$ ne segue che $D(\mathcal{A}, \mathcal{A}) > 0$ e quindi che $D(\mathcal{A}, \mathcal{B}) > \frac{1}{2} D(\mathcal{A}, \mathcal{A})$.

Dimostriamo ora che D è una pseudodistanza.

• A questo scopo è necessario dimostrare che valgono le proprietà di unicità delle pseudodistanze, simmetrica e transitiva. Nella dimostrazione scriveremo, per brevità, $\overline{AB}$ per $D(\mathcal{A}, \mathcal{B})$, $\overline{ab}$ per $d(a, b)$, $\underline{ab}_{AB}$ per $e_{A,B}(a, b)$. Scriveremo inoltre $d_{A,B}(a, b)$ per $e_{A,B}(a, b) + d(a, b)$.

La simmetria è banale dal fatto che, per ogni A_i, B_j si ha che $d_{A,B}(A_i, B_j) = d_{B,A}(B_j, A_i)$.

Per dimostrare la disuguaglianza triangolare si possono seguire i seguenti punti. Siano dati $\mathcal{A}, \mathcal{B}, \mathcal{C}$. Si avranno le seguenti espressioni:

$$\begin{aligned} \overline{AB} &= \frac{\sum_{a \in \mathcal{A}, b \in \mathcal{B}} d_{AB}(a, b)}{|\mathcal{A}| |\mathcal{B}|} \\ &> \frac{\sum_{b \in \mathcal{B}, c \in \mathcal{C}} \underline{bc}_{BC}}{|\mathcal{B}| |\mathcal{C}|} \\ \overline{BC} &= \frac{\sum_{b \in \mathcal{B}, c \in \mathcal{C}} d_{BC}(b, c)}{|\mathcal{B}| |\mathcal{C}|} \\ &> \frac{\sum_{a \in \mathcal{A}, c \in \mathcal{C}} \underline{ac}_{AC}}{|\mathcal{A}| |\mathcal{C}|} \end{aligned}$$

e si voglia dimostrare che $\overline{AB} + \overline{BC} \geq \overline{AC}$.

Si vede immediatamente che

$$\overline{AB} + \overline{BC} = \frac{\sum_{a \in \mathcal{A}, b \in \mathcal{B}} d_{AB}(a, b) + \sum_{b \in \mathcal{B}, c \in \mathcal{C}} d_{BC}(b, c)}{|\mathcal{A}| |\mathcal{B}| |\mathcal{C}|} = \frac{\sum_{a \in \mathcal{A}, b \in \mathcal{B}, c \in \mathcal{C}} (d_{AB}(a, b) + d_{BC}(b, c))}{|\mathcal{A}| |\mathcal{B}| |\mathcal{C}|} \text{ e che}$$

$$\overline{AC} = \frac{\sum_{a \in \mathcal{A}, c \in \mathcal{C}} d_{AC}(a, c)}{|\mathcal{A}| |\mathcal{C}|}. \text{ Basterà quindi confrontare i due numeratori, cioè confrontare}$$

$$\sum_{a \in \mathcal{A}, b \in \mathcal{B}, c \in \mathcal{C}} (d_{AB}(a, b) + d_{BC}(b, c)) \text{ con } \sum_{a \in \mathcal{A}, b \in \mathcal{B}, c \in \mathcal{C}} d_{AC}(a, c).$$

Nel seguito useremo la notazione d_{abc} per $e_{A,B} \dot{Y}a, b \dot{p} \dot{6} d \dot{Y}a, b \dot{p} + e_{B,C} \dot{Y}b, c \dot{p} \dot{6} d \dot{Y}b, c \dot{p}$, ovvero per il termine generico di $\overline{AB} + \overline{BC}$, e d_{abc}^v per $e_{A,C} \dot{Y}a, c \dot{p} \dot{6} d \dot{Y}a, c \dot{p}$, ovvero per il termine generico di $\overline{AC}$. Basterà quindi confrontare d_{abc} con d_{abc}^v .

Si considerino i singoli termini delle sommatorie, ovvero $d_{abc} = \underline{ab}_{AB} \dot{6} \overline{ab} + \underline{bc}_{BC} \dot{6} \overline{bc}$ e $d_{abc}^v = \underline{ac}_{AC} \dot{6} \overline{ac}$. Questi termini soddisfano la condizione data tutte le volte che $\underline{ac}_{AC} \dot{6} \overline{ac} \geq \min\{\underline{ab}_{AB}, \underline{bc}_{BC}\}$, ovvero tutte le volte che $\underline{ac}_{AC} = \frac{1}{2}$ oppure $\underline{ab}_{AB} = \underline{bc}_{BC} = 1$. Infatti in questi casi la disuguaglianza cercata deriva dalla disuguaglianza triangolare $\overline{ab} + \overline{bc} \geq \overline{ac}$.

Si hanno quindi tre casi in cui i termini non soddisfano immediatamente la disuguaglianza data: 1) quando si ha $\underline{ab}_{AB} = \frac{1}{2}$, $\underline{bc}_{BC} = 1$ e $\underline{ac}_{AC} = 1$; 2) quando si ha $\underline{ab}_{AB} = 1$, $\underline{bc}_{BC} = \frac{1}{2}$ e $\underline{ac}_{AC} = 1$; 3) quando si ha $\underline{ab}_{AB} = \frac{1}{2}$, $\underline{bc}_{BC} = \frac{1}{2}$ e $\underline{ac}_{AC} = 1$. Trattiamo separatamente i tre casi.

1. Consideriamo il primo caso. Dal fatto che $\underline{ab}_{AB} = \frac{1}{2}$ ne segue che $\dot{A}a, b \dot{A} \hat{O} A \vee B$. Questo caso si può spezzare in due: 1a) sia $a \dot{5} C$ e $c \dot{5} B$ e quindi $\underline{ca}_{BC} = \frac{1}{2}$ e, da $\underline{ac}_{AC} = 1$, anche $c \dot{6} A$; 1b) sia $\underline{ca}_{BC} = 1$. Consideriamo innanzitutto il caso 1b); nella sommatoria $\overline{AB} + \overline{BC}$ compaiono i termini $d_{abc} + d_{bac} = \underline{ab}_{AB} \dot{6} \overline{ab} + \underline{bc}_{BC} \dot{6} \overline{bc} + \underline{ba}_{AB} \dot{6} \overline{ba} + \underline{ac}_{BC} \dot{6} \overline{ac}$ e in $\overline{AC}$ compaiono i termini $d_{abc}^v + d_{bac}^v = \underline{ac}_{AC} \dot{6} \overline{ac} + \underline{bc}_{AC} \dot{6} \overline{bc}$. Per le condizioni date si ha che $d_{abc} + d_{bac} = \frac{1}{2} \overline{ab} + \overline{bc} + \frac{1}{2} \overline{ba} + \overline{ac} = \overline{ab} + \overline{bc} + \overline{ac} \geq \overline{ac} + \overline{bc} = d_{abc}^v + d_{bac}^v$. Nel caso 1a) si considerino invece i tre termini

$$\begin{aligned} d_{abc} + d_{bac} + d_{bca} &= \underline{ab}_{AB} \dot{6} \overline{ab} + \underline{bc}_{BC} \dot{6} \overline{bc} + \underline{ba}_{AB} \dot{6} \overline{ba} + \underline{ac}_{BC} \dot{6} \overline{ac} + \underline{bc}_{AB} \dot{6} \overline{bc} + \underline{ca}_{BC} \dot{6} \overline{ca} \\ &= \frac{1}{2} \overline{ab} + \overline{bc} + \frac{1}{2} \overline{ba} + \frac{1}{2} \overline{ac} + \overline{bc} + \frac{1}{2} \overline{ca} = \overline{ab} + 2\overline{bc} + \overline{ac} \\ &\quad + \underline{ac}_{AC} \dot{6} \overline{ac} + \underline{bc}_{AC} \dot{6} \overline{bc} + \underline{ba}_{AC} \dot{6} \overline{ba} = d_{abc}^v + d_{bac}^v + d_{bca}^v \end{aligned}$$

da cui la tesi.

2. Il secondo caso può essere dimostrato analogamente al primo caso.

3. Per il terzo caso, si considerino i tre termini

$$d_{abc} + d_{bac} + d_{acb} = \underline{ab}_{AB} \dot{6} \overline{ab} + \underline{bc}_{BC} \dot{6} \overline{bc} + \underline{ba}_{AB} \dot{6} \overline{ba} + \underline{ac}_{BC} \dot{6} \overline{ac} + \underline{ac}_{AB} \dot{6} \overline{ac} + \underline{cb}_{BC} \dot{6} \overline{cb}.$$

Nell'ipotesi peggiore è $a \dot{5} C$ e $c \dot{5} B$, da cui $\underline{ac}_{BC} = \frac{1}{2}$ e $\underline{ac}_{AB} = \frac{1}{2}$; si ha quindi

$$\begin{aligned} d_{abc} + d_{bac} + d_{acb} &= \frac{1}{2} \overline{ab} + \frac{1}{2} \overline{bc} + \frac{1}{2} \overline{ba} + \frac{1}{2} \overline{ac} + \frac{1}{2} \overline{ac} + \frac{1}{2} \overline{cb} = \overline{ab} + \overline{bc} + \overline{ac} \\ &\quad + \underline{ac}_{AC} \dot{6} \overline{ac} + \underline{bc}_{AC} \dot{6} \overline{bc} + \underline{ab}_{AC} \dot{6} \overline{ab} = d_{abc}^v + d_{bac}^v + d_{bca}^v \end{aligned}$$

da cui segue la disuguaglianza triangolare per la funzione D .

Per l'unicità delle pseudodistanze, si può dimostrare la condizione equivalente che se $A \dot{6} B$ si ha $D \dot{Y}A, B \dot{p} > D \dot{Y}A, A \dot{p}$. Considereremo separatamente il caso in cui A e B sono disgiunti, quello in cui $A \hat{O} B$, quello in cui $A \hat{O} B$ e infine il caso generale.

1. La tesi per A e B disgiunti deriva direttamente dal lemma precedente.
2. Sia $A \hat{O} B$. Si ponga $N = B \dot{?} A$, tale quindi che $N \vee A = 2$ e $|A| + |N| = |B|$. In base al lemma si avrà quindi:

$$\begin{aligned} D \dot{Y}A, A \dot{p} &= \frac{|A|}{|B|} \dot{6} D \dot{Y}A, A \dot{p} + \frac{|N|}{|B|} \dot{6} D \dot{Y}A, A \dot{p} \\ &< \frac{|A|}{|B|} \dot{6} D \dot{Y}A, A \dot{p} + \frac{|N|}{|B|} \dot{6} D \dot{Y}A, N \dot{p} \\ &= \frac{|A| \dot{6} > \frac{1}{2} d \dot{Y}A_i, A_j \dot{p}}{|A|^2 \dot{6} |B|} + \frac{|N| \dot{6} > d \dot{Y}A_i, N_k \dot{p}}{|A| \dot{6} |B| \dot{6} |N|} \\ &= \frac{A_i A_j \dot{5} A}{|A| \dot{6} |B|} + \frac{A_i \dot{5} A, N_k \dot{5} N}{|A| \dot{6} |B|} \\ &= D \dot{Y}A, B \dot{p} \end{aligned}$$

3. Sia invece $B \hat{O} A$. Si ponga anche $M = A \dot{?} B$ tale quindi che $M \vee B = 2$ e $|B| + |M| = |A|$.

Si avranno quindi le seguenti identità:

$$\begin{aligned}
 D\dot{Y}A, A\mathfrak{p} &= \frac{\frac{>}{B_i, B_j 5B} \frac{1}{2} d\dot{Y}B_i, B_j\mathfrak{p} + \frac{>}{B_i 5B, M_j 5M} \frac{1}{2} d\dot{Y}B_i, M_j\mathfrak{p} + \frac{>}{M_i 5M, B_j 5B} \frac{1}{2} d\dot{Y}M_i, B_j\mathfrak{p} + \frac{>}{M_i, M_j 5M} \frac{1}{2} d\dot{Y}M_i, M_j\mathfrak{p}}{|A|^2} \\
 &= \frac{|B|^2 \mathfrak{6} D\dot{Y}B, B\mathfrak{p} + \frac{1}{2} |B| \mathfrak{6} |M| \mathfrak{6} D\dot{Y}B, M\mathfrak{p} + \frac{1}{2} |M| \mathfrak{6} |B| \mathfrak{6} D\dot{Y}M, B\mathfrak{p} + |M|^2 \mathfrak{6} D\dot{Y}M, M\mathfrak{p}}{|A|^2} \\
 &= \frac{|B|^2 \mathfrak{6} D\dot{Y}B, B\mathfrak{p} + |B| \mathfrak{6} |M| \mathfrak{6} D\dot{Y}B, M\mathfrak{p} + |M|^2 \mathfrak{6} D\dot{Y}M, M\mathfrak{p}}{|A|^2}
 \end{aligned}$$

Dal precedente punto 1 si può osservare che $D\dot{Y}M, M\mathfrak{p} < D\dot{Y}B, M\mathfrak{p}$ in quanto **B** e **M** sono disgiunti. Aggiungendo inoltre il termine $|B| \mathfrak{6} |M| \mathfrak{6} D\dot{Y}B, B\mathfrak{p}$ si otterrà quindi:

$$\begin{aligned}
 D\dot{Y}A, A\mathfrak{p} &< \frac{|B|^2 \mathfrak{6} D\dot{Y}B, B\mathfrak{p} + |B| \mathfrak{6} |M| \mathfrak{6} D\dot{Y}B, M\mathfrak{p} + |M|^2 \mathfrak{6} D\dot{Y}B, M\mathfrak{p} + |B| \mathfrak{6} |M| \mathfrak{6} D\dot{Y}B, B\mathfrak{p}}{|A|^2} \\
 &= \frac{|B| \mathfrak{6} \dot{Y}B + |M| \mathfrak{p} \mathfrak{6} D\dot{Y}B, B\mathfrak{p} + |M| \mathfrak{6} \dot{Y}B + |M| \mathfrak{p} \mathfrak{6} D\dot{Y}B, M\mathfrak{p}}{|A|^2} \\
 &= \frac{|B| \mathfrak{6} |A| \mathfrak{6} D\dot{Y}B, B\mathfrak{p} + |M| \mathfrak{6} |A| \mathfrak{6} D\dot{Y}B, M\mathfrak{p}}{|A|^2} = \frac{|B|^2 \mathfrak{6} D\dot{Y}B, B\mathfrak{p} + |B| \mathfrak{6} |M| \mathfrak{6} D\dot{Y}M, B\mathfrak{p}}{|A| \mathfrak{6} |B|} \\
 &= \frac{\frac{>}{B_i, B_j 5B} \frac{1}{2} d\dot{Y}B_i, B_j\mathfrak{p} + \frac{>}{M_i 5M, B_j 5B} d\dot{Y}M_i, B_j\mathfrak{p}}{|A| \mathfrak{6} |B|} = D\dot{Y}A, B\mathfrak{p}
 \end{aligned}$$

4. Si consideri infine il caso generale con $A \vee B = C$, $M = A ? B$, $N = B ? A$, **C**, **M**, **N** disgiunti, $|C| + |M| = |A|$ e $|C| + |N| = |B|$. Si avrà:

$$\begin{aligned}
 D\dot{Y}A, A\mathfrak{p} &= \frac{\frac{>}{A_i, A_j 5A} \frac{1}{2} d\dot{Y}A_i, A_j\mathfrak{p}}{|A|^2} = \frac{|B| \mathfrak{6} \frac{>}{A_i, A_j 5A} \frac{1}{2} d\dot{Y}A_i, A_j\mathfrak{p}}{|A|^2 \mathfrak{6} |B|} \\
 &= \frac{|N| \mathfrak{6} \frac{>}{A_i, A_j 5A} \frac{1}{2} d\dot{Y}A_i, A_j\mathfrak{p} + |C| \mathfrak{6} \frac{>}{A_i, A_j 5A} \frac{1}{2} d\dot{Y}A_i, A_j\mathfrak{p}}{|A|^2 \mathfrak{6} |B|}
 \end{aligned}$$

I due addendi del denominatore possono essere analizzati separatamente. Si avrà, per il primo:

$$\begin{aligned}
 |N| \mathfrak{6} \frac{>}{A_i, A_j 5A} \frac{1}{2} d\dot{Y}A_i, A_j\mathfrak{p} &= \frac{1}{2} \mathfrak{6} \frac{>}{A_i, A_j 5A, N_k 5N} d\dot{Y}A_i, A_j\mathfrak{p} \\
 &= \frac{1}{2} \mathfrak{6} \frac{>}{A_i, A_j 5A, N_k 5N} \dot{Y}d\dot{Y}A_i, N_k\mathfrak{p} + d\dot{Y}N_k, A_j\mathfrak{p} \\
 &= \frac{|A|}{2} \mathfrak{6} \left(\frac{>}{A_i 5A, N_k 5N} d\dot{Y}A_i, N_k\mathfrak{p} + \frac{>}{A_j 5A, N_k 5N} d\dot{Y}N_k, A_j\mathfrak{p} \right) \\
 &= |A| \mathfrak{6} \frac{>}{A_i 5A, N_k 5N} d\dot{Y}A_i, N_k\mathfrak{p}
 \end{aligned}$$

Per il secondo verrà considerata solo la sommatoria; con una disuguaglianza simile a quella utilizzata al precedente punto 3 si avrà:

$$\begin{aligned}
& \geq \frac{1}{2} d'_{A_i, A_j} = \geq \frac{1}{2} d'_{M_i, M_j} + \geq \frac{1}{2} d'_{M_i, C_j} + \geq \frac{1}{2} d'_{C_i, M_j} + \geq \frac{1}{2} d'_{C_i, C_j} \\
& = |M|^2 \cdot d'_{M, M} + \frac{1}{2} |M| \cdot |C| \cdot d'_{M, C} + \frac{1}{2} |C| \cdot |M| \cdot d'_{C, M} + |C|^2 \cdot d'_{C, C} \\
& = |M|^2 \cdot d'_{M, M} + |M| \cdot |C| \cdot d'_{M, C} + |C|^2 \cdot d'_{C, C} \\
& < |M|^2 \cdot d'_{M, C} + |M| \cdot |C| \cdot d'_{M, C} + |C|^2 \cdot d'_{C, C} + |M| \cdot |C| \cdot d'_{C, C} \\
& = |M| + |C| \cdot d'_{M, C} + |C| \cdot d'_{C, C} = |A| \cdot d'_{M, C} + |C| \cdot d'_{C, C} \\
& = |A| \cdot \left(\geq d'_{M_i, C_j} + \geq \frac{1}{2} d'_{C_i, C_j} \right)
\end{aligned}$$

Ne seguirà quindi:

$$\begin{aligned}
D'_{A, A} &= \frac{|N| \cdot \geq \frac{1}{2} d'_{A_i, A_j} + |C| \cdot \geq \frac{1}{2} d'_{A_i, A_j}}{|A|^2 \cdot |B|} \\
&< \frac{|A| \cdot \geq d'_{A_i, N_k} + |A| \cdot \left(\geq d'_{M_i, C_j} + \geq \frac{1}{2} d'_{C_i, C_j} \right)}{|A|^2 \cdot |B|} \\
&= \frac{\geq d'_{A_i, N_j} + \geq d'_{M_i, C_j} + \geq \frac{1}{2} d'_{C_i, C_j}}{|A| \cdot |B|} = D'_{A, B}
\end{aligned}$$

n

In genere la funzione D descritta non è una distanza; esistono infatti elementi A tali che $D'_{A, A} \neq 0$. Si consideri per esempio $A = \hat{a}\hat{a}, \hat{b}\hat{a}, \hat{a}\hat{a}, \hat{c}\hat{a}$. Si avrà (considerando sempre la distanza d e la misura W citate precedentemente) che

$$\begin{aligned}
D'_{A, A} &= \frac{d_{A, A} \cdot \hat{a}\hat{a}, \hat{b}\hat{a}, \hat{a}\hat{a}, \hat{b}\hat{a} + d_{A, A} \cdot \hat{a}\hat{a}, \hat{b}\hat{a}, \hat{a}\hat{a}, \hat{c}\hat{a} + d_{A, A} \cdot \hat{a}\hat{a}, \hat{c}\hat{a}, \hat{a}\hat{a}, \hat{b}\hat{a} + d_{A, A} \cdot \hat{a}\hat{a}, \hat{c}\hat{a}, \hat{a}\hat{a}, \hat{c}\hat{a}}{2 \times 2} \\
&= \frac{0 + \frac{1}{2} \cdot d'_{\hat{a}\hat{a}, \hat{b}\hat{a}, \hat{a}\hat{a}, \hat{c}\hat{a}} + \frac{1}{2} \cdot d'_{\hat{a}\hat{a}, \hat{c}\hat{a}, \hat{a}\hat{a}, \hat{b}\hat{a}} + 0}{4} \\
&= \frac{\left(\frac{W_{\hat{a}\hat{a}, \hat{c}\hat{a}}}{W_{\hat{a}\hat{a}, \hat{b}, \hat{c}\hat{a}}} \right)}{4} = \frac{1}{6}
\end{aligned}$$

Si possono avere peraltro elementi A tali che $D'_{A, A} = 0$. Per esempio si ha immediatamente che la funzione D descritta è una distanza considerando i soli simboli atomici, in quanto, posto $A = \hat{a}\hat{a}\hat{a}$, è $D'_{A, A} = \frac{1}{2} d'_{A, A} = 0$.

In effetti va evidenziato che, nel caso in cui tutti gli elementi di A siano finiti, $D'_{A, A} = 0$ se e solo se A è atomico. Basta infatti notare che, se A non è atomico, esistono almeno due elementi distinti $A_1, A_2 \in A$. Per la definizione di distanza è $d'_{A_1, A_2} > 0$, e quindi $D'_{A, A} \geq \frac{d'_{A_1, A_2}}{|A|^2} > 0$. È peraltro necessario che A_1 e A_2 non siano entrambi infiniti: infatti si possono trovare insiemi di cardinalità infinita tali che $d'_{A_1, A_2} = 0$ con $A_1 \in A_2$.

Proprietà legate alla d e alla W definite

2001-04-23

Si consideri in particolare la distanza tra insiemi d definita come $d'_{A, B} = k \cdot \frac{W_{A \# B}}{W_{A \cup B}}$, dove W è una misura qualunque. In questo caso si avranno alcune proprietà aggiuntive per la pseudodistanza D .

Innanzitutto si può dimostrare che $0 \leq D'_{A, B} \leq k$. Infatti, da $0 \leq d'_{A, B} \leq k$ e da

$e_{A,B} d(A_i, B_j) \geq 1$ segue che $D(A, B) = \frac{|A| \cdot |B|}{|A| \cdot |B|} \geq \frac{|A| \cdot |B|}{|A| \cdot |B|} = k$. In particolare si può dimostrare la condizione più restrittiva che $0 \leq D(A, B) \leq \frac{k}{2}$.

Siano disponibili un numero infinito di simboli $a_1, a_2, \dots, a_i, \dots$. Per ogni possibile P e per ogni cardinalità m , si possono allora trovare due simboli A_m e B_m di cardinalità m , tali che $0 < D(A_m, B_m) < P$.

O

La tesi può essere dimostrata inizialmente per simboli atomici. Sia dato $n \in \mathbb{N}$ tale che $\frac{2}{n+1} < P$. Sia $A_1 = \{a\}$ e $B_1 = \{b\}$ dove $A = \{a_{x_1}, \dots, a_{x_n}, a\}$ e $B = \{a_{x_1}, \dots, a_{x_n}, b\}$ con $a \neq b$. Sarà $D(A_1, B_1) = D(A, B) = \frac{2}{n+1} < P$.

Per estendere la dimostrazione al caso generale, dimostreremo in particolare che per ogni $m > 1$ esistono simboli non atomici A_m per i quali si ha $0 < D(A_m, A_m) < P$. Sia $A_m = \{a_{x_1}, \dots, a_{x_n}, a\}$ tale che $A_1 = \{a_{x_1}, \dots, a_{x_n}, a\}$, $A_2 = \{a_{x_1}, \dots, a_{x_n}, a_2\}$, $\dots$, $A_m = \{a_{x_1}, \dots, a_{x_n}, a_m\}$ dove tutti gli $x_1, \dots, x_n, a_1, \dots, a_m$ sono fra di loro diversi. Si ha che

$$d(A_i, A_j) = \begin{cases} 0 & \text{se } i = j \\ \frac{W(A_i, A_j)}{W(A_i, A_j)} = \frac{2}{n} & \text{se } i \neq j \end{cases}$$

da cui

$$\begin{aligned} D(A, A) &= \frac{\sum_{i=1, \dots, m} d(A_i, A_j)}{m^2} = \frac{\sum_{i=j} d(A_i, A_j) + \sum_{i \neq j} d(A_i, A_j)}{m^2} \\ &= \frac{0 + \sum_{i \neq j} \frac{2}{n}}{m^2} = \frac{2}{n} \cdot \frac{m(m-1)}{m^2} < \frac{2}{n} \end{aligned}$$

n

In maniera analoga si può dimostrare che per ogni possibile P e per ogni cardinalità m , si può trovare un simbolo A_m di cardinalità m , tale che $0 \leq D(A_m, A_m) < P$. In particolare, se $m = 1$ si ha $0 = D(A_1, A_1)$, mentre per $m \geq 2$ è sempre $0 < D(A_m, A_m)$. Per dimostrare la tesi basta considerare, per $m = 2$, un insieme $A_2 = \{a_{x_1}, a_{x_2}\}$ con $A_1 = \{a_{x_1}, a\}$ e $A_2 = \{a_{x_1}, b\}$, con $\frac{2}{n+1} < P$.

In presenza di un numero infinito di simboli $a_1, a_2, \dots, a_i, \dots$, si può inoltre dimostrare anche che per ogni possibile $P < k$ e per ogni cardinalità m , si possono allora trovare due simboli A_m e B_m di cardinalità m , tali che $P < D(A_m, B_m) < k$.

Si può infine dimostrare anche che, per ogni possibile $P < \frac{k}{2}$ e per ogni cardinalità $m > 1$, si può trovare un simbolo A_m di cardinalità m , tale che $P < D(A_m, A_m) < \frac{k}{2}$.

Estensione al caso infinito 2001-05-18

Come osservato precedentemente, si possono analizzare a parte i problemi relativi all'utilizzo di insiemi A e B infiniti. In questo paragrafo considereremo in particolare il caso in cui le cardinalità di A e B siano numerabili, ponendo quindi $A = \{a_{x_1}, \dots, a_{x_n}, a\}$ e $B = \{a_{x_1}, \dots, a_{x_n}, b\}$. Affinché questo sia possibile è richiesto che il supporto dei nostri insiemi sia esso stesso almeno numerabile, in quanto, altrimenti, non vi potrebbe essere un numero infinito di insiemi A_m distinti. Si ha infatti che $|A| \leq |P^U| = 2^{|U|}$.

Innanzitutto va ridefinita la pseudodistanza D , in quanto la definizione precedente è ben data solo per cardinalità finite. Per arrivare a una definizione alternativa considereremo i seguenti punti.

Innanzitutto, operando su cardinalità numerabile, possiamo dare un ordinamento totale su tutti i sottoinsiemi di U . Si possono allora ordinare gli elementi $A_1, \dots, A_m, \dots$ e $B_1, \dots, B_m, \dots$ in

maniera coerente all'ordinamento dato. Definiamo allora la successione $\mathbf{A}_0, \cup, \mathbf{A}_m, \cup$ nel modo seguente: per ogni i , si definisce $\mathbf{A}_i = \hat{\mathbf{a}}_{A_1, \cup, A_i, \hat{\mathbf{a}}}$, considerando i primi i insiemi A_i nell'ordine precedentemente dato. Si noti che la definizione data è ben posta anche nel caso in cui $\mathbf{A}$ abbia cardinalità finita; in questo caso, per ogni $m \geq |\mathbf{A}|$ si avrà $\mathbf{A}_m = \mathbf{A}$.

Definiremo allora la pseudodistanza D nel modo seguente:

$$D(\mathbf{A}, \mathbf{B}) = \lim_{m \rightarrow \infty} D(\mathbf{A}_m, \mathbf{B}_m)$$

Si noti che, per $\mathbf{A}$ e $\mathbf{B}$ numerabili, il limite qui indicato è uguale alla pseudodistanza precedentemente definita. In particolare, considerando la distanza $d(\mathbf{A}, \mathbf{B}) = k \frac{|\mathbf{A} \setminus \mathbf{B}|}{|\mathbf{A} \cup \mathbf{B}|}$, si avrà che anche nel caso infinito $0 \leq D(\mathbf{A}, \mathbf{B}) \leq k$.

Una interessante differenza rispetto al caso finito è che, nel caso infinito, esistono elementi $\mathbf{A}$ non atomici tali che $D(\mathbf{A}, \mathbf{A}) = 0$. Come precedentemente osservato questo è innanzitutto possibile considerando insiemi $A_1 \otimes A_2$ tali che $d(A_1, A_2) = 0$; in questo caso, posto $\mathbf{A} = \hat{\mathbf{a}}_{A_1, A_2, \hat{\mathbf{a}}}$, sarà $D(\mathbf{A}, \mathbf{A}) = 0$. Per trovare due insiemi cosiffatti, basta considerare i due insiemi infiniti $A_1 = \hat{\mathbf{a}}_{a, a_2, \cup, a_m, \cup, \hat{\mathbf{a}}}$ e $A_2 = \hat{\mathbf{a}}_{b, a_2, \cup, a_m, \cup, \hat{\mathbf{a}}}$, tali cioè che $A_1 \setminus A_2 = \hat{\mathbf{a}}_{a, b, \hat{\mathbf{a}}}$ e $|A_1 \cap A_2| = K$. Si possono comunque trovare, nel caso infinito, anche situazioni in cui $D(\mathbf{A}, \mathbf{A}) = 0$ pur essendo in genere $d(A_i, A_j) > 0$. Come esempio si consideri $\mathbf{A} = \hat{\mathbf{a}}_{A, A_1, A_2, \cup, A_i, \cup, \hat{\mathbf{a}}}$ con $A = \hat{\mathbf{a}}_{a, \hat{\mathbf{a}}}$, $A_1 = \hat{\mathbf{a}}_{a, a_2, \cup, a_i, \cup, \hat{\mathbf{a}}}$, $A_2 = \hat{\mathbf{a}}_{a_1, a, \cup, a_i, \cup, \hat{\mathbf{a}}}$, e, in generale, $A_i = \hat{\mathbf{a}}_{a_1, a_2, \cup, a_{i-1}, a, a_{i+1}, \cup, \hat{\mathbf{a}}}$.

Applicando la definizione si ha che

$$\begin{aligned} D(\mathbf{A}, \mathbf{A}) &= \lim_{m \rightarrow \infty} \frac{\frac{1}{2} \sum_{i=1, \cup, m-1} d(\mathbf{A}, A_i) + \frac{1}{2} \sum_{j=1, \cup, m-1} d(A_j, \mathbf{A})}{m^2} \\ &= \lim_{m \rightarrow \infty} \frac{\sum_{i=1, \cup, m-1} d(A_i, A_i)}{m^2} = \lim_{m \rightarrow \infty} \frac{\sum_{i=1, \cup, m-1} 0}{m^2} = 0 \end{aligned}$$

In genere, se $\mathbf{A} = \hat{\mathbf{a}}_{A_1, \cup, A_i, \cup, \hat{\mathbf{a}}}$ è di cardinalità infinita, con le cardinalità degli A_i limitate superiormente da n , ovvero con $|A_i| < n$, si ha che $D(\mathbf{A}, \mathbf{A}) > 0$. Infatti si ha che $d(A_i, A_j) > \frac{k}{n}$ e quindi che $D(\mathbf{A}, \mathbf{A}) > \lim_{m \rightarrow \infty} \frac{m^2 \frac{k}{n}}{m^2} = \frac{k}{n}$. Anche in questo caso si possono però trovare, dato P , degli elementi $\mathbf{A}$ del tipo descritto tali che $D(\mathbf{A}, \mathbf{A}) < P$. Si scelga, come limite superiore delle cardinalità degli A_i un valore n tale che $\frac{1}{n} < P$. Si ponga inoltre, per ogni i , $A_i = \hat{\mathbf{a}}_{x_1, \cup, x_{n-2}, a_i, \cup, \hat{\mathbf{a}}}$ con tutti gli $x_1, \cup, x_{n-2}, a_1, \cup, a_i, \cup$ diversi tra di loro. Si ha allora che, scelti comunque A_i, A_j , si ha $d(A_i, A_j) = \begin{cases} 0 & \text{se } i = j \\ \frac{2}{n} & \text{se } i \otimes j \end{cases}$. Si avrà allora che

$$D(\mathbf{A}, \mathbf{A}) = \lim_{m \rightarrow \infty} \frac{\sum_{i,j} d(A_i, A_j)}{m^2} = \lim_{m \rightarrow \infty} \frac{\frac{1}{n} m^2}{m^2} = \frac{1}{n} < P.$$

problemi

Quali sono le proprietà delle distanze che non valgono per le pseudodistanze?
Quali di queste proprietà sono critiche nell'applicazione che ci interessa?

Note

1. Metriche su un piano? ovvero $d \in \mathbb{R}^+$ piuttosto che $d \in \mathbb{R}, 1 \leq d \leq \infty$

2. Utilizzabilità su reti neurali? Ovvero, esistono dei meccanismi di parallelizzazione delle espressioni descritte? (a un primo livello, quello delle macchine vettoriali, sembra di sì; è però utile individuare se esistono meccanismi più evoluti)
3. Definizione su insiemi infiniti di simboli? sì, nei due esempi seguenti (che fanno entrambi riferimento alla funzione $W^0 : U \rightarrow \mathbb{R}^+$):

¾ quando si ha che $\sum_{x \in U} W^0(x) \leq 5$

¾ quando si accettano solo le informazioni A tali che $|A| < K$;

¾ forse estendendo la definizione di W nel modo seguente: $d(A, B) = k \cdot \frac{W(A \# B)}{W(A \cup B)}$ per ogni A e B tali che $|A| < K$ e $|B| < K$; altrimenti $d(A, B) = \lim_{C^v \in AWB} d(A \cup C^v, B \cup C^v)$ con C^v

finiti?

attenzione, creando la successione dei C^v in maniere diverse si ottengono limiti diversi: si deve definire con precisione come va creata la sequenza dei C^v ; per esempio, sia

$A = \{0, 2, 4, 6, 8, 10, 12\}$, $u \in A$ e $B = \{0, 3, 6, 9, 12\}$, $u \in B$; si ha

$A \cup B = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $u \in A \cup B$ e $A \cap B = \{0, 6, 12\}$, $u \in A \cap B$; posti

$C_0 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_1 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_2 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_3 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_4 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_5 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$ si ottiene $d(A, B) = 3/4$;

posti invece

$C_0 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_1 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_2 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_3 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_4 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$, $C_5 = \{0, 2, 3, 4, 6, 8, 9, 10, 12\}$

Si noti che il tutto sembra funzionare, nei termini così descritti, per U numerabile.

4. La definizione data fa riferimento a insiemi A di simboli elementari $a_1, \dots, a_n$. In realtà, uno dei casi che interesserebbe gestire è per esempio quello delle parole, che sono *sequenze ordinate* di simboli elementari. Come gestire l'eventuale concetto di ordine? Un caso ancor più generale è quello in cui le informazioni possono essere strutturate in maniera complessa.

Sequenziabilità delle informazioni

L'ultima nota fa riferimento a informazioni descritte in sequenza (o comunque strutturate). In questi casi bisogna distinguere $A = \{a, b, c\}$ da $B = \{b, a, c\}$.

Da ragionarci sopra.

Ipotesi non funzionante: porre $A = \{a, b, c\} = \{a, a, a, b, a, a, b, c, a, a\}$ e

$B = \{b, a, c\} = \{a, a, a, b, a, a, b, c, a, a\}$; si avrebbe quindi (posto $W^0(a) = 1$ e $k = 1$,

$d(A, B) = \frac{W(A \# B)}{W(A \cup B)} = \frac{2}{5}$. Il problema consiste nel fatto che se $A = \{a, b, c\}$ e $C_1 = \{x, b, c\}$ si ha che $d(A, C_1) = \frac{6}{7}$, mentre se $A = \{a, b, c\}$ e $C_3 = \{a, b, x\}$ si ha che $d(A, C_3) = \frac{1}{5}$.

Tentativo #2001.06.04#

Definizione standard

$$D(A, B) = \frac{\sum_{i,j} d(A_i, B_j)}{|A| \cdot |B|}$$

1. iperpiano: $\{f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n)\} = \{0\}$ con f lineare. Si avrà cioè $f_1(x_1, \dots, x_n) = a_{11}x_1 + \dots + a_{1n}x_n$.
2. iperretta: $\{f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n)\} = \{0\}$ con f_1 e f_2 lineari. Si avrà cioè $f_1(x_1, \dots, x_n) = a_{11}x_1 + \dots + a_{1n}x_n$ e $f_2(x_1, \dots, x_n) = a_{21}x_1 + \dots + a_{2n}x_n$.
3. qualunque sottospazio di dimensione $m < n$ sarà definito come $\{f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n)\} = \{0\}$. Si avrà quindi una matrice dei coefficienti delle funzioni $C = [c_{ij}]$, con $c_{ij} = 0$ per $i > m$.
4. In particolare, un punto, di dimensione 0, sarà individuato da una matrice con tutte le linee indipendenti, e quindi corrisponderà a una matrice con determinante non nullo.
5. La dimensione del sottospazio è la differenza tra il numero totale di linee e il numero di linee linearmente indipendenti nella matrice.
6. A questo punto, come si calcola la distanza tra sottospazi?

Nel caso di parole, la distanza ha una semantica. Quali sono le distanze tra "abc", "abx", "ab", "cab", "xyz"? quali di queste distanze è minima? Ci aspettiamo che $d(abc, abc) < d(abc, abx) < d(abc, xyz)$, ma per il resto? È più importante l'ordine delle lettere o quali lettere compaiono? (nel primo caso si ha che $d(abc, abx) < d(abc, cab)$, mentre nel secondo si avrebbe il viceversa) È più importante la presenza delle stesse lettere o l'assenza di discrepanze nelle lettere? (nel primo caso si ha che $d(abc, ab) < d(abc, abx)$, mentre nel secondo si avrebbe il viceversa)

Inoltre il tutto va integrato con gli aspetti relativi agli algoritmi di ordinamento tipo Unicode, in cui l'ordinamento va fatto rispetto ad almeno tre chiavi (lettera, diacritico, caso: $a < b$, $a < \grave{a}$, $a < A$). In altri termini: differenze di caratteri contano più di differenze di diacritici e differenze di diacritici contano più di differenze di caso.

Metriche su un piano

tentativo; definire la distanza nel modo seguente:

$$d(A, B) = \frac{|A \# B| + \frac{|A| + |B|}{2}}{|A| + |B|}$$

va però dimostrato che d così definito sia una distanza. in particolare va dimostrato che

$$d(A, B) + d(B, C) \geq d(A, C)$$

Equivale a dover dimostrare che

$$|A \# B| + \frac{|A| + |B|}{2} + |B \# C| + \frac{|B| + |C|}{2} \geq |A \# C| + \frac{|A| + |C|}{2}$$

La differenza simmetrica è una distanza, infatti

$$|A \# B| + |B \# C| \geq |A \# C|;$$

dato che anche $\frac{|A \# B|}{|A| + |B|}$ è una distanza.

Non è possibile dimostrare la tesi cercata partendo dalla relazione $a + b \leq c$ e $x + y \leq z$ implicano $ax + by \leq cz$. Infatti si ha che $cz = (a + b)(x + y) = ax + by + bx + ay$.