Bollettino Unione Matematica Italiana

Sezione A – La Matematica nella Società e nella Cultura

Tiziana Pacelli

Similarity, distanze ed informazione incompleta

Bollettino dell'Unione Matematica Italiana, Serie 8, Vol. **10-A**—*La Matematica nella Società e nella Cultura* (2007), n.2, p. 307–310.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=BUMI_2007_8_10A_2_307_0>

Bollettino dell'Unione Matematica Italiana, Unione Matematica Italiana, 2007.

Bollettino U. M. I.
La Matematica nella Società e nella Cultura
Serie VIII, Vol. X-A, Agosto 2007, 307-310

Similarity, distanze ed informazione incompleta

TIZIANA PACELLI

1. – Introduzione.

L'ipotesi per cui sia possibile considerare la nozione di *regione* come primitiva al posto di quella di *punto* risale ad alcune interessanti ricerche di A. N. Whitehead ([7]). Successivamente le idee di Whitehead sono state riprese da vari autori ed elaborate da diversi punti di vista. Ad esempio in ([4]) viene proposto un approccio di tipo metrico in cui sono assunte come primitive le nozioni di regione, inclusione tra regioni, distanza tra regioni e diametro di una regione.

La tesi parte dal convincimento che le ricerche sulla *geometria senza punti*, anche se nascono nell'ambito di problemi dei fondamenti della matematica, hanno valenze che vanno oltre tale problematica. Ad esempio, con riferimento all'approccio metrico, osserviamo che, in molti contesti riguardanti processi di elaborazione di informazione e di classificazione, è fondamentale confrontare gli oggetti investigati e fornire una misura di quanto questi oggetti siano "dissimili" o, se si vuole, di quanto siano "simili". In tal caso, questi oggetti sono visti come "punti" e per misurare la loro "dissimilarità" si fa ricorso all'usuale concetto di "distanza", formalmente rappresentatata mediante la nozione di spazio metrico. In alternativa, spesso viene utilizzata la nozione logica di "similarità" che dualizza quella di distanza e che nasce nell'ambito degli studi sulla logica a più valori. Il problema che sorge è se tali nozioni siano adeguate in quei contesti in cui non siamo in grado di ottenere informazione completa sugli oggetti considerati. I pezzi di informazione disponibili possono essere visti come regioni di uno spazio piuttosto che punti. Ciò conduce alla esigenza di definire *distanze generalizzate ed approssimate* che siano adeguate a rappresentare un tale tipo di situazione. Dualmente, sul versante logico, si pone il problema di definire *similarità generalizzate*. In entrambi i casi, se le regioni vengono ad essere reinterpretate come "*pezzi incompleti di informazione*", il diametro di una regione rappresenta una misura dell'incompletezza di tale informazione. Minore è il diametro, più l'informazione è completa. I punti (che hanno diametro zero) rappresentano un informazione completa.

2. – Dualità tra distanza e similarity su regioni.

Tra il concetto di distanza e quello di *similarity* c'è una dualità facilmente comprensibile: quando si confrontano due oggetti in accordo alle loro proprietà, è possibile usare sia una misura di quanto siano "dissimili" sia una misura di quanto siano "simili". La seconda nozione sembra più adatta della prima in quei contesti in cui il confronto tra oggetti avviene in termini delle proprietà da essi soddisfatte. Un primo argomento che si affronta nella tesi è di estendere tale dualità per stabilire un collegamento tra la geometria senza punti e l'approccio categoriale alla teoria dei fuzzy set proposto da H—hle in [6]. In particolare, riferendosi alla definizione di *spazi metrici senza punti*, si introduce la nozione di *spazi ultrametrici senza punti*([2]):

DEFINIZIONE 1. – *Uno spazio ultrametrico senza punti (spazio PU) è una struttura* $(R, \leq, \delta, |\ |)$, *in cui* $(R, \leq\ )$ *è un insieme ordinato,* $\delta : R \times R \rightarrow [0, \infty)$ *è una funzione decrescente,* $|\ | : R \rightarrow [0, \infty]$ *è una funzione crescente e, per ogni* $x, y, z \in R$, *valgono i seguenti assiomi:*

a1) $\delta(x, x) = 0$,
a2) $\delta(x, y) = \delta(y, x)$,
a3) $\delta(x, y) \leq \delta(x, z) \vee \delta(z, y) \vee |z|$. (disuguaglianza triangolare generalizzata forte)

In ambito logico la nostra attenzione si focalizza sulla nozione di semisimilarity con la t-norma di Gödel, chiamata G-semisimilarity.

DEFINIZIONE 2. – *Sia* $\wedge$ *la t-norma di Gödel. Una G-semisimilarity è una fuzzy relazione E su R tale che*

e1) $E(x, y) = E(y, x)$, *(simmetria)*
e2) $E(x, z) \wedge E(z, y) \leq E(x, y)$. *(transitività)*

Viene mostrata l'esistenza di una connessione diretta tra strutture con G-semisimilarity e spazi ultrametrici senza punti.

TEOREMA 1. – *Sia E una G-semisimilarity. Definiamo* $\|_E : R \rightarrow [0, 1]$ *ponendo* $|x|_E = 1 - E(x, x)$ *e* $\delta_E : R \times R \rightarrow [0, 1]$ *ponendo* $\delta_E(x, y) = 0$ *se* $E(x, y) = E(x, x) \wedge E(y, y)$ *e* $\delta_E(x, y) = 1 - E(x, y)$ *se* $E(x, y) < E(x, x) \wedge E(y, y)$, *per ogni* $x, y \in R$. *Allora* $(R, \delta_E, \|_E)$ *è uno spazio PU. Viceversa, sia* $(R, \delta, \|)$ *uno spazio PU e definiamo* $E_{\delta,\|} : R \times R \rightarrow [0, 1]$ *ponendo* $E_{\delta,\|}(x, y) = 1 - (\delta(x, y) \vee |x| \vee |y|)$. *Allora* $E_{\delta,\|}$ *è una G-semisimilarity.*

Infine, dopo aver organizzato la classe degli spazi ultrametrici in una categoria, vengono definiti due funtori per collegare tale categoria con quella di Höhle.

3. – Distanze approssimate.

Completando quanto proposto in [4], nella tesi si definisce una *distanza approssimata* intesa come funzione che ad ogni coppia di oggetti associa un intervallo ([3]). In accordo con le idee dell'*Interval Analysis*, tale intervallo è interpretato come un constraint sulla distanza effettiva e si riduce ad un numero relae solo quando si ha una conoscenza completa dei punti.

DEFINIZIONE 3. – *Uno spazio semimetrico a valori intervallo (spazio ISM) è una struttura* $(R, \leq, \Delta)$, *in cui* $(R, \leq)$ *è un insieme ordinato,* $\Delta : R \times R \to I(\mathbf{R}_0^+)$ *è una funzione a valori in un intervallo di reali positivi, tale che, per ogni* $x, y, z \in R$, *valgono i seguenti assiomi:*

A1) $\Delta(x, x) \cdot [0, 1] = \Delta(x, x)$,
A2) $\Delta(x, y) = \Delta(y, x)$,
A3) $\Delta(x, y) - \Delta(z, z) \leq_I \Delta(x, z) + \Delta(z, y)$,
A4) $\Delta(x, y)) - \Delta(x, y) \leq_I \Delta(x, x) + \Delta(y, y)$,
A5) $x \leq x', y \leq y' \Rightarrow \Delta(x, y) \subseteq \Delta(x', y')$.

Da notare che le operazioni tra intervalli e la relazione $\leq_I$ sono quelle usualmente definite nell'Interval Analysis. Chiamiamo *regioni* gli elementi di R e chiamiamo *funzione peso* la funzione $p : R \to \boldsymbol{R}_0^+$ definita da

$$p(x) = \pi_2(\Delta(x, x)),$$

per ogni $x \in R$, dove π_2 è la proiezione seconda. Possiamo interpretare un elemento in R come un "pezzo incompleto di informazione", $\Delta(x, y)$ come una misura approssimata di quanto i due oggetti descritti dai pezzi di informazione x e y siano lontani e il peso $p(x)$ come una misura della completezza di x.vcViene anche esaminato come la nozione di distanza approssimata possa essere applicata in alcuni campi dell'informatica. In particolare, si fa riferimento alla teoria degli insiemi fuzzy dove vengono proposte due definizioni di distanza-intervallo, mediante la nozione di taglio e mediante quella di ipografo. Inoltre, si definisce una distanza-intervallo tra rough set ed, infine, come applicazione al clustering, viene esaminato un algoritmo basato sulle distanze intervallo tra clusters.

4. – Similarity e sistemi esperti.

Nella tesi si descrive, infine, come la nozione di similarità permetta di definire un metodo per la costruzione di sistemi esperti di natura probabilistica ([1]). Si fa riferimento all'idea, proposta in ([5]), secondo cui il nostro "grado di credenza" circa una proprietà relativa ad un nuovo caso da esaminare derivi dalla nostra esperienza circa una classe di casi passati "simili" al caso che si sta esaminando. Si considerano "simili" due casi che verificano le stesse proprietà "osservabili" e si propone una misura di similarità. Nel fare questo si usa un formalismo che estende quello della *formal concept analysis*. Più precisamente, si accetta la possibile presenza di proprietà vaghe e ci si riferisce ad un noto teorema sulle similarità.

TEOREMA 2. – *Sia* $\otimes$ *la t-norma di Lukasiewicz e sia* $\leftrightarrow_\otimes$ *l'operazione di bire-siduo ad essa associata. Dato un insieme S e una famiglia* s_i *di fuzzy sottoinsiemi di S, la fuzzy relazione* $E{:}S \times S \to [0,1]$ *definita ponendo* $E(x,y) = \otimes s_i(x) \leftrightarrow_\otimes s_i(y)$, *è una* $\otimes$*-fuzzy similarity.*

Da notare che, se si interpreta s_i come l'estensione di una proprietà vaga, $E(x,y)$ è la valutazione, in una logica a più valori, dell'affermazione " *x è simile a y se x ed y soddisfano le stesse proprietà*". Facendo riferimento a tale misura di similarità, il sistema esperto fornisce la probabilità con cui il nuovo caso che si sta esaminando verifica una proprietà "non osservabile". Inizialmente, si suppone che il sistema esperto abbia nel suo database casi passati tutti simili al caso attuale. Poi, è possibile ottenere informazione sul caso attuale mediante una sequenza $a_1, ..., a_n$ di appropriate domande relative a proprietà osservabili. Poniamo $T_0 = \emptyset$ e, data una nuova domanda a_i poniamo $T_i = T_{i-1} \cup (a_i, \lambda_i)$, dove $\lambda_i \in [0,1]$ è il grado con cui il caso attuale verifica la proprietà a_i. Se l'informazione ottenuta è ritenuta "sufficente", il sistema esperto fornisce la probabilità cercata, che in altre parole rappresenta la percentuale dei casi passati simili al caso attuale che nel passato hanno verificato la proprietà non osservabile. Se, invece, non si ha ancora informazione sufficiente allora si continua con la strategia delle domande per ottenere ulteriore informazione.

RIFERIMENTI BIBLIOGRAFICI

[1] COPPOLA C., GERLA G. e PACELLI T. *Similarities for Crisp and Fuzzy Probabilistic Expert Systems,* inviato a Studies in Fuzziness and Soft Computing, Springer.

[2] COPPOLA C., Gerla G. e Pacelli T. *Point-free Ultrametric Spaces and the Category of Fuzzy Subsets,* Proceedings of the Tenth International Conference IPMU 2004 **1** (2004,), 503-510.

[3] COPPOLA C., PACELLI T. *Approximate distances, pointless geometry and incomplete information,* Fuzzy Sets and Systems **157,17** (2006), 2371-2383.

[4] GERLA G. *Pointless Metric Spaces,* Journal of Symbolic Logic **55**, 1 (1990), 207-219.

[5] GERLA G. *The probability that Tweety is able to fly,* International Journal of Intelligent Systems **9** (1994,), 403-409

[6] Höhle U. *Presheaves over GL-monoid,* in Nonclassical Logics and Their Applications to Fuzzy Subsets Kluwer (1995), 127-157.

[7] Whitehead A.N., Process and Reality, Univ. Press, Cambridge (1929).

Dipartimento di Matematica e Informatica, Università degli Studi di Salerno
e-mail: tpacelli@unisa.it
Dottorato in Matematica
(sede amministrativa: Università degli Studi di Salerno) - Ciclo XVIII.
Direttore di ricerca: Prof. Giangiacomo Gerla, Università degli Studi di Salerno