ESTRUCTURA DE LAS SIMILARIDADES

Ramón Gonzalez del Campo [1] **Luis Garmendia** [2] **Jordi Recasens** [3]

[1] SIC. Facultad de Informática, rgonzale@estad.ucm.es
[2] DISIA. Facultad de Informática. UCM, lgarmend@fdi.ucm.es
[3] Universitat Politècnica de Catalunya. j.recasens@upc.edu

Resumen

En este artículo se define formalmente el concepto de estructura de similaridad, se cuenta el número de estructuras de similaridades hasta dimensión 5, se propone una nomenclatura y un algoritmo que asigna una estructura a cada similaridad.

Palabras Clave: Relación borrosa, Min-transitividad, Similaridad.

1 INTRODUCCIÓN

Las relaciones borrosas de similaridad permiten dar grados de parecido entre los elementos de un universo y encuentran numerosas aplicaciones en métodos de clustering y de inferencia borrosa. Sus alfa-cortes son relaciones de equivalencia clásicas, por lo que ofrecen numerosas "rejillas" o particiones de un universo. Por otro lado son Min-preórdenes [1], por lo que proporcionan conclusiones según Tarski cuando se utilizan para hacer inferencia mediante la regla composicional de inferencia.

En el estudio de estas importantes relaciones, nos damos cuanta de que muchas de ellas se parecen mucho y otras no. Nos parece muy definir lo que vamos a llamar "estructura" de similaridad, que nos ayude a entenderlas y a clasificarlas según a qué estructura pertenezcan.

Este artículo comienza definiendo el concepto de estructura como clases de equivalencia de una relación de equivalencia RE, que relaciona dos similaridades si tienen la misma estructura. Para definir RE definimos otras dos relaciones de equivalencia, Rπ y RT, que relaciona dos similaridades si se pueden transformar una en otra mediante una permutación y mediante una relación monótona estricta.

Hemos estudiado y contado las estructuras de las similaridades hasta la dimensión 5

2 PRELIMINARES

Sea $E=\{e_1,...,e_n\}$ un conjunto finito. Una relación difusa es una transformación $R:E \times E \rightarrow [0,1]$. El grado de relación para los elementos e_i y e_j, $R(e_i,e_j)$, se denota como r_{ij}.

Definición 2.1: Un operador *binario* $T:[0,1] \times [0,1] \rightarrow [0,1]$ es una t-norma o norma triangular [4,2] si satisface:

- $T(1,x)=x$
- $T(y,x)=T(x,y)$
- $T(x,T(y,z))=T(T(x,y),z)$
- Si $x \leq x'$ y $\leq y'$ entonces $T(x,y) \leq T(x',y')$

Definición 2.2: Dada una t-norma T, una relación difusa $R:E \times E \rightarrow [0,1]$ es T-transitiva si $T(R(a,b),R(a,c)) \leq R(a,c)$ $\forall\, a,b,c \in E$. Esto es, si $T(r_{ik},r_{kj}) \leq r_{ij}$ $\forall\, i,j,k:1..n$.

Definición 2.3: Dada una relación difusa, R, es una Proximidad si safisface:

- Reflexiva: $r_{ii}=1$
- Simétrica: $r_{ij}=r_{ji}$ $\forall\, 1 \leq i, j \leq n$

Si además R es una relación T-transitiva, se denomina T-indistinguibilidad [6,7]. Si la T-norma es el mínimo, a esta T-indistinguibilidad se le llama Similaridad [8].

3 RELACIONES ENTRE RELACIONES BORROSAS

Definición 3.1: Sean A y B dos matrices de dimensión n, están relacionadas por R_T si y solo si $\exists$ T tal que $T: A \rightarrow B$, $T(a_{ij})=b_{ij}$ tal que es monótona estricta.

Lema 3.1: R_T es una relación de equivalencia.

Demostración:

- Es reflexiva con T igual a la función identidad.
- Es simétrica: si $A\ R_T\ B$ entonces $B\ R_T^{-1} A$.
- Es transitiva: si $A\ R_T\ B$ y $B\ R_{T'}\ C$ entonces: $A\ R_{T' \cdot T}\ C$.

El número de clases de equivalencia, S^n/R_T están contados en [5].

Lema 3.2: Dada una matriz de una similaridad *A*, si permutamos las filas y las correspondientes columnas, denotado por $B=P_\pi(A)$, obtenemos una matriz que representa a una similaridad.
Demostración:
Una permutación, $\pi=\{\pi_i\}\ \forall\ i:1..n$, consiste en reetiquetar los índices de la siguiente forma: $i \rightarrow \pi_i$.

- Es reflexiva: $b_{ii}=a_{\pi i\ \pi i}=1$
- Es simétrica: $b_{ij}=a_{\pi i\ \pi j}=a_{\pi j\ \pi i}=b_{ji}$
- Es min-transitiva: $min(b_{ik},b_{kj}\leq b_{ij})$ ya que $min(a_{\pi i \pi k},a_{\pi k\ \pi j}\})\leq a_{\pi i\ \pi j}$

Definición 3.2: Sean *A* y *B* dos matrices de dimensión *n*, están relacionadas por R_π si y solo si existe una permutación π tal que $B=P_\pi(A)$.

Lema 3.3: R_π es una relación de equivalencia.
Demostración:

- Es reflexiva: A R_π A pues $A=P_\pi(A)$ con $\pi=(1..n)$.
- Es simétrica: A R_πB $\leftrightarrow$ $B=P_\pi(A)$ $\leftrightarrow$ $P_\pi(B)=P_\pi(P_\pi(A))$ puesto que: $A=P_\pi(P_\pi(A))$.
- Es transitiva: A Rπ B, B Rπ C $\rightarrow$ $A=P_\pi(B),B=P_\pi(C)$ $\rightarrow$ $A=P_\pi(P_{\pi'}(C))$ $\rightarrow$ $A=P_{\pi\cdot\pi'}(C)$.

La siguiente relación de equivalencia indica si dos similaridades tienen la misma "estructura".

Lema 3.4: R_E es una relación de equivalencia.
Demostración:

- Es reflexiva: AR_EA tomando $\pi=(1..n)$ ya que $AR_\pi A$ y $A\ R_{Id}\,A$.
- Es simétrica por ser R_T y R_π simétricas.
- Es transitiva por ser R_T y R_π transitivas.

Definición 3.4: Sea S^n el conjunto de similaridades de dimensión *n*. Cada elemento de partes de $R_E(S^n/R_E)$ es una estructura de similaridad de dimensión *n*.

4 CONTEO DE ESTRUCTURAS DE SIMILARIDAD

Para investigar el número de estructuras de similaridad nos basamos en el siguiente lema:

Lema 4.1(Teorema de Lee [3]): Sean C y D dos relaciones borrosas t y:

$$E(t;C,D)=\begin{pmatrix} C & F^T \\ F & D \end{pmatrix}$$

Si C y D son similaridades entonces $E(t;C,D)=E$ también es una similaridad $\forall t\in [0,min(min(C),min(D))]$.

Para cada matriz, S^n, que representa una similaridad de dimensión *n* existe una descomposición:

$$S^n=P_\pi(E(t;C_{n1xn1},D_{n2xn2}))$$

Para todas las similaridades de dimensión n investigamos el número de estructuras diferentes a partir de las descomposiciones posibles: $E(t;C,D)$.

Casos:

n=2:
Solamente existe una estructura:

$$\begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$$

n=3:
Las descomposiciones con ($n_1=1,n_2=2$) y ($n_1=2,n_2=1$) representan a la misma estructura bajo la permutación $\pi=(3,2,1)$.

$$E(t;C,D)=\begin{pmatrix} \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} & t \\ t & 1 \end{pmatrix}$$

donde $a\geq t$. Por lo tanto, existen 2 estructuras:

$$\begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix} \text{ y } \begin{pmatrix} 1 & a & t \\ a & 1 & t \\ t & t & 1 \end{pmatrix}$$

donde $a>t$.

n=4:
- Las descomposiciones con $(n_1=1,n_2=3)$ y $(n_1=3,n_2=1)$ representan la misma estructura bajo la permutación $\pi=(4,2,3,1)$. Las similaridades de dimensión 4 son de dos posibles formas:
- Descomposición con $(n_1=3,n_2=1)$:

$$\begin{pmatrix} \begin{pmatrix} 1 & a & b \\ a & 1 & b \\ b & b & 1 \end{pmatrix} & \begin{matrix} t \\ t \\ t \end{matrix} \\ \begin{matrix} t & t & t \end{matrix} & 1 \end{pmatrix}$$

con $a\geq b\geq t$. Sustituyendo cada desigualdad por una desigualdad estricta o una igualdad obtenemos 4 posibles estructuras.
- Descomposición con $(n_1=2,n_2=2)$:

$$\begin{pmatrix} \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} & \begin{matrix} t & t \\ t & t \end{matrix} \\ \begin{matrix} t & t \\ t & t \end{matrix} & \begin{pmatrix} 1 & b \\ b & 1 \end{pmatrix} \end{pmatrix}$$

donde $a\geq t$ y $b\geq t$ $(a\in C$ y $b\in D)$.

Los casos con $b=t$ ya se han tenido en cuenta en la descomposición anterior. Tenemos 2 estructuras adicionales.

- En total, para $n=4$ existen 6 estructuras.

n=5:
Las descomposiciones con $(n_1=1,n_2=4)$ y $(n_1=4,n_2=1)$ representan la misma estructura bajo la permutación $\pi=(5,2,3,4,1)$.
Las descomposiciones con $(n_1=2,n_2=3)$: sean $a\in C$ y $b,\ c\in D$. Entonces se verifica:
Combinaciones posibles de valores de *a*,*b*,*c*:

- $a \ge b \ge c \ge t \rightarrow 2^3$ casos distintos
- $b \ge a \ge c \ge t \rightarrow 2^2$ casos distintos (porque *a*=*b* ya se ha tratado)
- $b \ge c \ge a \ge t \rightarrow 2$ casos distintos adicionales (porque *a*=*c* ya se ha tratado y si *a*=*t* obtenemos una estructura ya tratada en la descomposición anterior).

- En total, tenemos 22 estructuras distintas para *n=5*.

4.1 HACIA UNA NOMENCLATURA DE LAS ESTRUCTURAS

Sea *S* cualquier similaridad de dimensión *n*.
Consideremos la descomposición:

$$E(t;C,D)=\begin{pmatrix} C & F^T \\ F & D \end{pmatrix}$$

Sea *z* el número de elementos de *F* y F^T. Se verifica:

$$n^2=n_1^2+n_2^2+2n_1xn_2$$
$$z=2n_1xn_2$$

Sea $I=\{1..n_1+n_2-1\}$.

Denotamos cada estructura mediante la expresión: $S^{(n,z)}_{\{qi\}\{pi\}}$ donde:

- *n* es la dimensión, *z* es el número de elementos de las submatrices *F* y F^T.

- q_i es el número de veces que aparece el valor *i*. Para los elementos de *C* se cumple: $\sum_{i=1} q_i \le n_1^2-n_1$. El valor del puente, q_t, es el primer valor para el que se verifica $\sum_{i=1} q_i > n_1^2-n_1$. Para los elementos de D se cumple:

$$\sum_{i=qt+1} q_i \le n_2^2-n_2$$

Sea $\{p_i\}$ la lista de los valores distintos de *S*, además: $p_j \neq 1$, tal que si $k<j$ la magnitud del valor *k* es mayor que la magnitud del valor *j*. Si $\{p_i\}=\{1..k\}$ con $k \le n_1+n_2-1$ la omitimos.

Ejemplos:

- $\begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$ se denota por $S^{(2,2)}$.

- $\begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$ y $\begin{pmatrix} 1 & a & t \\ a & 1 & t \\ t & t & 1 \end{pmatrix}$ se denotan por $S^{(3,4)}_{\{6\}}$ y $S^{(3,4)}_{\{2,4\}}$ respectivamente.

- $\begin{pmatrix} 1 & a & b & t \\ a & 1 & b & t \\ b & b & 1 & t \\ t & t & t & 1 \end{pmatrix}$ con $a>b>t$ se denota $S^{(4,8)}_{\{2,4,8\}}$.

Si $a>b=t$ entonces: $S^{(4,8)}_{\{2,12\}}$

- $\begin{pmatrix} \begin{pmatrix} 1 & a & b \\ a & 1 & b \\ b & b & 1 \end{pmatrix} & t \\ t & \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix} \end{pmatrix}$ con $a>c>b=t$ se denota $S^{(5,12)}_{\{2,16,2\}\{1,3,2\}}$.

Este es el ejemplo más pequeño en el que podemos apreciar que la relación entre los valores de las similaridades *C* y *D* es relevante. En la matriz anterior si $c>a$, por ejemplo, no existe una función *T* monótona estricta que con la misma matriz en la que se dé $c<a$. Son estructuras diferentes: $S^{(5,12)}_{\{2,16,2\}\{3,1,2\}}$ y $S^{(5,12)}_{\{2,16,2\}\{1,3,2\}}$.

4.2 ALGORITMO DE CLASIFICACIÓN DE SIMILARIDADES

Podemos clasificar fácilmente cualquier similaridad de dimensión *n*, *S*, en alguna estructura de dimensión *n* con la que esté R_E relacionada, mediante el siguiente algoritmo.

Pasos:

- Sea $_{bi0\,j0}=t$ el mínimo valor.
- Sean $I = \{i \text{ tal que } i \in N \text{ y } b_{ij0}=t\}$, y $I' = N \setminus I$.
- Sea el vector π tal que: $\pi=I \cup I'$.
- Para cada $p,q \in \{1..n\}$ hacer: $a_{pq}=b_{\pi[p]\ \pi[q]}$.
- $z=2card(I)x\ card(I')$
- Sea $\{c_i\}$ la lista de las magnitudes de los valores de *C* y $\{d_i\}$ la lista de valores de *D*, ambos ordenados decrecientemente.
- Sea $\{a_j\}$ la lista de valores de *S* ordenados decrecientemente por su magnitud.
- Sea la lista $\{nc_i\}$ tal que nc_i es el número de veces que aparece el valor *i* en *C* y la lista $\{nd_i\}$ tal que nd_i es el número de veces que aparece el valor *j* en *D*.
- Para cada $p,q \in I$ hacer: si $a_{pq}=c_i$ incrementa nc_i.
- Para cada $p,q \in I'$ hacer: si $a_{pq}=d_i$ incrementa nd_i.
- Pertenece a $S^{n,z}_{\{\{nci\},z,\{ndi\}\}\{ai\}}$

Ejemplo:
Sea la similaridad, S, con $n=4$:

$$\begin{pmatrix} 1 & 0{,}4 & 0{,}7 & 0{,}4 \\ 0{,}4 & 1 & 0{,}4 & 0{,}6 \\ 0{,}7 & 0{,}4 & 1 & 0{,}4 \\ 0{,}4 & 0{,}6 & 0{,}4 & 1 \end{pmatrix}$$

Pasos:
- El mínimo valor es: $b_{12}=0{,}4$.
- $I=\{1,3\}$, $I'=\{2,4\}$.
- $\pi=\{1,3,2,4\}$.
- Obtenemos:

$$\begin{pmatrix} 1 & 0{,}7 & 0{,}4 & 0{,}4 \\ 0{,}7 & 1 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 1 & 0{,}6 \\ 0{,}4 & 0{,}4 & 0{,}6 & 1 \end{pmatrix}$$

Esta similaridad se clasifica en $S^{4,8}_{\{2,8,2\}\{1,2\}}$.

5 ÁRBOLES JERÁRQUICOS Y ESTRUCTURAS

Una de las razones del interés en el estudio de las similaridades es que generan árboles de particiones indexados. En efecto, para cada entrada de a de la similaridad S definida en el universo X, la relación $\approx$ crisp: $x_i \approx x_j$ si y sólo si $S(x_i,x_j) \geq a$ es una partición P_a de X y si $a \geq b$, entonces P_a es un refinamiento de P_b.

Definición 5.1:
Un árbol es un grafo sin ciclos.
Los nodos con una sola arista se llaman hojas.
La distancia entre dos nodos es el número de aristas que los separan.

Definición 5.2:
Un árbol jerárquico es un árbol con un nodo privilegiado *A* tal que:
- La distancia d entre *A* y las hojas del árbol es constante.
- El cardinal de los nodos a distancia *k* de *A* es menor que el cardinal de los nodos a distancia *k+1* de *A* para todo $k=1,\ldots,d$.

Es inmediato comprobar los siguientes resultados.

Proposición 5.3:
Dos relaciones de similaridad generan el mismo árbol (no indexado) si y sólo si son equivalentes por R_T.

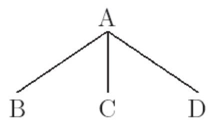
Proposición 5.4:
Dos relaciones de similaridad generan el mismo árbol indexado salvo permutaciones de sus ramas si y sólo si son equivalentes por R_π.

Proposición 5.5:
Dos relaciones de similaridad generan el mismo árbol (no indexado) salvo permutaciones de sus ramas si y sólo si son equivalentes por R_E.

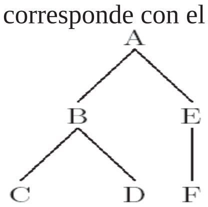
La búsqueda de estructuras de similaridades es pues equivalente a hallar todos los árboles jerárquicos salvo permutaciones de sus ramas.

En la representación de las estructuras mediante árboles jerárquicos:
- cada nodo representa una similaridad,
- cada nodo hoja representa un objeto del universo *X*.
- las aristas que unen nodos de dos niveles consecutivos representan el valor de la relación de los objetos y/o similaridades.

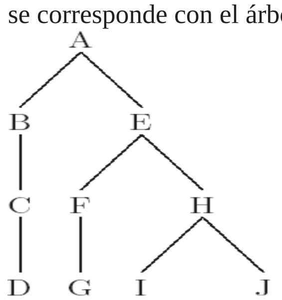
Ejemplos:
- La similaridad $S^{(3,4)}_{\{6\}}$: $\begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$ se corresponde con el árbol:



- La similaridad $S^{(3,4)}_{\{2,4\}}$: $\begin{pmatrix} 1 & a & b \\ a & 1 & b \\ b & b & 1 \end{pmatrix}$ se corresponde con el árbol:



- La similaridad $S^{(4,8)}_{\{2,4,8\}}$: $\begin{pmatrix} 1 & a & b & t \\ a & 1 & b & t \\ b & b & 1 & t \\ t & t & t & 1 \end{pmatrix}$ se corresponde con el árbol:



6 CONCLUSIONES Y TRABAJO FUTURO

Una vez hemos definido formalmente el concepto de estructura, hemos encontrado una notación adecuada para representar cada estructura y hemos encontrado una algoritmo que decide a que estructura pertenece una similaridad, permitiendo una clasificación de las similaridades según su estructura y un mejor entendimiento de las mismas que podría ser considerado para estudiar el comportamiento de las similaridades en distintas aplicaciones según su forma.

Seguimos buscando una fórmula, seguramente recursiva, que nos cuente el número de estructuras de similaridades de dimensión mayor a 5, y un algoritmo que genere un representante de dichas estructuras para cualquier dimensión, que podría acercarnos una manera de representación, a una mejor comprensión y como generador universal de todas las similaridades finitas.

Referencias

[1] J. Elorza y P. Burillo. On the relation between fuzzy preorders and fuzzy consequence operators. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 7(3):219{234, 1999}.

[2] J. Jacas y J. Recasens. Fuzzy t-transitive relations: eigenvectors and generators. Fuzzy Sets Syst., 72(2):147{154, 1995}.

[3] Hsuan-Shih Lee. An optimal algorithm for computing the max-min transitive closure of a fuzzy similarity matrix. Fuzzy Sets and Systems, 123(1):129{136, 2001}.

[4] Sklar A. Schweizer, B. Probabilistic Metric Spaces. Dover Publications, 1984.

[5] Riera T. How similarity matrices are? Stochastica, 2(4), 1978.

[6] Valverde L. Trillas, E. An inquiry into indistinguishability operators, in Aspects of Vagueness. Reidel Pubs, 1984.

[7] L. Valverde. On the structure of f-indistinguishability operators. Fuzzy Sets and Systems 17, pages 313{328, 1985}.

[8] L. A. Zadeh. Similarity relations and fuzzy orderings. Inform. Sci. 3, pages 177{200, 1971}.