Functional dependencies over domains with similarities
A Comparative Survey II

Lucie Urbanová

DATA ANALYSIS AND MODELING LAB

Palacky University, Olomouc, Czech Republic

esf european social fund in the czech republic

EUROPEAN UNION

MINISTRY OF EDUCATION, YOUTH AND SPORTS

OP Education for Competitiveness

INVESTMENTS IN EDUCATION DEVELOPMENT

Overview

Similarity-based approaches IX

2005: B. K. Tyagi, A. Sharfuddin, R. N. Dutta, D. K. Tayal

- Data: Possibilistic distribution
- Ranks from [0,1]
- Fuzzy equality for domain values: For all $u, v, w \in D_y$:
  - $E(u, v) = 1$ iff $u = v$
  - $E(u, v) = 1$ iff $E(v, u) = 1$
  - $\min\{E(u, v), E(v, w)\} \leq E(u, w)$

Fuzzy Functional dependencies

$\pi_{AB}(\mathcal{D})(r, r') = \bigvee\{\mathcal{D}(r'')|r'' \in Tupl(R)$ such that $r''(A) = r, r''(B) = r'\}$

FFD $A \Rightarrow B$, where $A, B \subseteq R$, is satisfied if $\forall r_1, r_2 \in D_A$ and $\forall r'_1, r'_2 \in D_B$:

$$\pi_{AB}(\mathcal{D})(r_1, r'_1) \wedge \pi_{AB}(\mathcal{D})(r_2, r'_2) \wedge E(r_1(A), r_2(A)) \leq E(r'_1(B), r'_2(B))$$

Non-similarity based approach

1991: Kiss - ordinary data, no similarity, ranks

$$\forall r_1, r_2 : ((D(r_1) \wedge D(r_2) \wedge r_1(A) = r_2(A)) \Rightarrow r_1(B) = r_2(B)).$$

Semantics:

- $\wedge, \forall$: the operator $\inf$
- $\vee, \exists$: the operator $\sup$
- $\Rightarrow$: Łukasiewicz implication
- $\neg a$: $1 - a$ for all $a \in [0, 1]$

Truth value to which the fuzzy relation $\mathcal{D}$ satisfies a given FD was given by the following

$$||A \Rightarrow B||_{\mathcal{D}} = 1 - \sup\{\inf(\mathcal{D}(r_1), \mathcal{D}(r_2)) | \, r_1(A) = r_2(A), r_1(B) \neq r_2(B))\}$$

Kiss used the following rule: $\neg(a \rightarrow_L b) = a \wedge \neg b$

Similarity-based approaches X

2009: Cordero et al: Fuzzy Attributes Tables

- Complete residuated lattice; Crisp data
- Ranks assigned to each attribute value ("degree of confidence or certainty")

$$\mathcal{D}\colon \prod_{y_i \in R} D_i \to L^Y$$

For each tuple r: Tuple of truth values $(\mathcal{D}(r))(y_i)_{y_i \in R}$.

- Similarity ($A \subseteq R$):

$$r_1 \approx_i r_2 = (\mathcal{D}(r_1))(y_i) \otimes (\mathcal{D}(r_2))(y_i) \to (r_1(y_i) \approx_i r_2(y_i))$$

$$(r_1(A) \approx_A r_2(A)) = \bigwedge_{y_i \in A} (r_1 \approx_i r_2)$$

A *fuzzy functional dependency* is an expression $A \overset{\theta}{\Rightarrow} B$ where $A, B \subseteq R$ and $\theta \in [0,1]$.

$$\theta \leq \bigwedge_{r_1, r_2 \in \prod_{y_i \in R} D_{y_i}} (r_1(A) \approx_A r_2(A)) \to (r_1(B) \approx_B r_2(B))$$

Similarity-based approaches XI

Belohlavek, Vychodil:

- Complete residuated lattice with hedge
- Crisp data
- Ranks:

$$\mathcal{D}\colon \prod_{y_i \in R} D_{y_i} \to L$$

- Similarity ($A, B \in L^R$):

$$r_1(A) \approx_{\mathcal{D}} r_2(A) = (\mathcal{D}(r_1) \otimes \mathcal{D}(r_2)) \to \bigwedge_{y \in R} (A(y) \to r_1(y) \approx_y r_2(y))$$

Similarity-based FD:

$$||A \Rightarrow B||_{\mathcal{D}} = \bigwedge_{r_1, r_2 \in \mathrm{Tupl}(R)} \Big((r_1(A) \approx_{\mathcal{D}} r_2(A))^* \to (r_1(B) \approx_D r_2(B)) \Big)$$

Similarity-based approaches: Comparison

$R = \{y_1, y_2\}$, $D_{y_1} = D_{y_2} = \{0.9, 0.8, 0.1\}$, $\approx_{y_1} = \approx_{y_2}$, FD: $||\{y_1\} \Rightarrow \{y_2\}||_{\mathcal{D}} = ?$

$\approx$	0.9	0.8	0.1
0.9	1	0.9	0.2
0.8	0.9	1	0.3
0.1	0.2	0.3	1

$\mathcal{D}$	y_1	y_2
	0.9	0.9
	0.9	0.8

$\mathcal{D}'$	y_1	y_2
0.3	0.9	0.9
0.2	0.9	0.1

Similarity-based approaches: Comparison

$R = \{y_1, y_2\}$, $D_{y_1} = D_{y_2} = \{0.9, 0.8, 0.1\}$, $\approx_{y_1} = \approx_{y_2}$, FD: $||\{y_1\} \Rightarrow \{y_2\}||_{\mathcal{D}} = ?$

$\approx$	0.9	0.8	0.1
0.9	1	0.9	0.2
0.8	0.9	1	0.3
0.1	0.2	0.3	1

$\mathcal{D}$	y_1	y_2
	0.9	0.9
	0.9	0.8

$\mathcal{D}'$	y_1	y_2
0.3	0.9	0.9
0.2	0.9	0.1

Classical FD	0	Not applicable (0)
Buckles, Petry: $\beta = 0.7$ $\beta * C(A[r_1, r_2]) \leq C(B[r_1, r_2])$	1	Not applicable (0)
Raju, Majumdar: $r_1(A) \approx_A r_2(A) \leq r_1(B) \approx_B r_2(B)$	0	0

Similarity-based approaches: Comparison

$R = \{y_1, y_2\}$, $D_{y_1} = D_{y_2} = \{0.9, 0.8, 0.1\}$, $\approx_{y_1} = \approx_{y_2}$, FD: $||\{y_1\} \Rightarrow \{y_2\}||_{\mathcal{D}} = ?$

$\approx$	0.9	0.8	0.1
0.9	1	0.9	0.2
0.8	0.9	1	0.3
0.1	0.2	0.3	1

$\mathcal{D}$	y_1	y_2
	0.9	0.9
	0.9	0.8

$\mathcal{D}'$	y_1	y_2
0.3	0.9	0.9
0.2	0.9	0.1

Prade, Testemale: $\lambda = 0.7$

$r_1(A) = r_2(A) \rightarrow (r_1(B) \approx_B r_2(B) \geq \lambda)$

1 — Not applicable (0)

Chen: $\theta = 0.7$

$(r_1(A) \approx_A r_2(A) \rightarrow_G r_1(B) \approx_B r_2(B)) \geq \theta$

1 — Not applicable (0)

Chen (2): $\theta = 0.7$

If: $r_1(A) = r_2(A) \rightarrow r_1(B) = r_2(B)$

Else:

$(r_1(A) \approx_A r_2(A) \rightarrow_G r_1(B) \approx_B r_2(B)) \geq \theta$

0 — Not applicable (0)

Similarity-based approaches: Comparison

$R = \{y_1, y_2\}$, $D_{y_1} = D_{y_2} = \{0.9, 0.8, 0.1\}$, $\approx_{y_1} = \approx_{y_2}$, FD: $||\{y_1\} \Rightarrow \{y_2\}||_{\mathcal{D}} = ?$

$\approx$	0.9	0.8	0.1
0.9	1	0.9	0.2
0.8	0.9	1	0.3
0.1	0.2	0.3	1

$\mathcal{D}$	y_1	y_2
	0.9	0.9
	0.9	0.8

$\mathcal{D}'$	y_1	y_2
0.3	0.9	0.9
0.2	0.9	0.1

Bhuniya, Niyogi: $\beta = 0.7$

$r_1(A) \approx_A r_2(A) \leq r_1(B) \approx_B r_2(B)$ or $(r_1(A) \approx_A r_2(A) - r_1(B) \approx_B r_2(B)) \leq 1 - \beta$

1

0

Cubero et al: $\alpha, \beta = 0.7$

$(r_1(A) \approx_A r_2(A) \geq \alpha) \rightarrow (r_1(B) \approx_B r_2(B) \geq \beta)$

1

Not applicable (0)

Similarity-based approaches: Comparison

$R = \{y_1, y_2\}$, $D_{y_1} = D_{y_2} = \{0.9, 0.8, 0.1\}$, $\approx_{y_1} = \approx_{y_2}$, FD: $||\{y_1\} \Rightarrow \{y_2\}||_{\mathcal{D}} = ?$

$\approx$	0.9	0.8	0.1
0.9	1	0.9	0.2
0.8	0.9	1	0.3
0.1	0.2	0.3	1

$\mathcal{D}$	y_1	y_2
	0.9	0.9
	0.9	0.8

$\mathcal{D}'$	y_1	y_2
0.3	0.9	0.9
0.2	0.9	0.1

Ben Yahia et al: $\lambda = 0.7$

$\min_{r_1, r_2} (r_1(A) \approx_A r_2(A)) \rightarrow_L (r_1(B) \approx_B r_2(B))$

0.9 | 0

Bosc, Pivert (Łukasiewicz)

$\forall r_1, r_2 \in \mathcal{D}$:

$r_1(A) \approx_A r_2(A) \rightarrow r_1(B) \approx_B r_2(B)$

0.9 | Not applicable (0.2)

Tyagi et al:

$\pi_{AB}(\mathcal{D})(r_1, r_1') \wedge \pi_{AB}(\mathcal{D})(r_2, r_2') \wedge$

$\wedge E(r_1(A), r_2(A)) \leq E(r_1'(B), r_2'(B))$

0 | 1

Similarity-based approaches: Comparison

$R = \{y_1, y_2\}$, $D_{y_1} = D_{y_2} = \{0.9, 0.8, 0.1\}$, $\approx_{y_1} = \approx_{y_2}$, FD: $||\{y_1\} \Rightarrow \{y_2\}||_{\mathcal{D}} = ?$

$\approx$	0.9	0.8	0.1
0.9	1	0.9	0.2
0.8	0.9	1	0.3
0.1	0.2	0.3	1

$\mathcal{D}$	y_1	y_2
	0.9	0.9
	0.9	0.8

$\mathcal{D}'$	y_1	y_2
0.3	0.9	0.9
0.2	0.9	0.1

Kiss:

$1 - \bigvee\{(\mathcal{D}(r_1) \wedge \mathcal{D}(r_2))|$
$r_1(A) = r_2(A), r_1(B) \neq r_2(B))\}$

0

0.8

Bel, Vych: (Łukasiewicz)

$\bigwedge_{r_1, r_2 \in \mathrm{Tupl}(R)} \Big((r_1(A) \approx_{\mathcal{D}} r_2(A))^* \rightarrow (r_1(B) \approx_D r_2(B))\Big)$

0.9

1

Similarity-based approaches: Comparison

Belohlavek, Vychodil: General approach
Similarity ($A, B \in L^R$):

$$r_1(A) \approx_{\mathcal{D}} r_2(A) = (\mathcal{D}(r_1) \otimes \mathcal{D}(r_2)) \to \bigwedge_{y \in R} (A(y) \to r_1(y) \approx_y r_2(y))$$

SBFD: $||A \Rightarrow B||_{\mathcal{D}} = \bigwedge_{r_1, r_2 \in \mathrm{Tupl}(R)} \Big((r_1(A) \approx_{\mathcal{D}} r_2(A))^* \to (r_1(B) \approx_D r_2(B)) \Big)$

- Particular choice of residuated implication
- Particular choice of the structure of truth degrees
- If Ranks are presented, but not employed in the definition of FD:
  Consider new $\mathcal{D}'$ as strong 0-cut of $\mathcal{D}$: $\mathcal{D}'(r) = 1$ iff $\mathcal{D} > 0$
- Additional parameters appears: $(r_1(A) \approx_A r_2(A) \geq \alpha) \to (r_1(B) \approx_B r_2(B) \geq \beta)$
  Using the fact that $A, B \in L^R$: $A(y) = \alpha$ for $y \in A$; $B(y) = \beta$ for $y \in B$
- Rank assigned to every tuple value: New domain $D_y \times L$ and new similarity :

$$\langle d_1, a_1 \rangle \approx_y \langle d_2, a_2 \rangle = (a_1 \otimes a_2) \to (d_1 \approx_y d_2)$$

Another approach

1997: Wei-Yi Liu

- Semantic proximity
- Attribute values are intervals

Semantic proximity between two fuzzy values f_1, f_2: $SP(f_1, f_2) \in [0, 1]$
For $f_1 = [a_1, b_1]$, $f_2 = [a_2, b_2]$, $g_1 = [c_1, d_1]$, $g_2 = [c_2, d_2]$:

$SP(f_1, f_2) = 1$ iff $a_1 = a_2 = b_1 = b_2$
$SP(f_1, f_2) = 0$ iff $f_1 \cap f_2 = \emptyset$
If $a_1 = a_2, b_1 = b_2, c_1 = c_2, d_1 = d_2$ and $|d_1 - c_1| > |b_1 - a_1|$ then $SP(f_1, f_2) \geq SP(g_1, g_2)$
If $|a_2 - b_2| = |a_1 - b_1|$ and $|f_1 \cup g_1| \geq |f_2 \cup g_1|$, then $SP(f_1, g_1) \geq SP(f_2, g_1)$

Then FD $A \Rightarrow B$ holds in a relation $\mathcal{D}$ iff

$$\forall r_1, r_2 \in \mathcal{D} : SP(r_1(A), r_2(A)) \leq SP(r_1(B), r_2(B))$$

Dang in 2004: Inference rules given by Liu are not sound

FFDs as Linguistic summaries

1998: Bosc, Lietard, Pivert: **Extended FD as a Basis for Linguistic Summaries**
For A, B single attributes and L_i, L_j linguistic label on A, B:
$(A, L_i) \to (B, L_j)$ holds iff for all $r \in \mathcal{D}$ we have

$$\mu_{L_i}(r(A)) \to_{R-G} \mu_{L_j}(r(B)),$$

$$\mu_{L_i}(r(A)) \leq \mu_{L_j}(r(B))$$

.
Meaning: The more $r(A)$ is L_i, the more $r(B)$ is L_j.
The *taller* the player the *higher* score in NBA.

In general: $(A_1, L_1), \ldots, (A_p, L_p) \to (B_1, L_{p+1}), \ldots, (B_q, L_q)$ is valid iff for all $r \in \mathcal{D}$:

$$\min\{\mu_{L_1}(r(A_1)), \ldots, \mu_{L_p}(r(A_p))\} \leq \min\{\mu_{L_{p+1}}(r(B_1)), \ldots, \mu_{L_q}(r(B_q))\}$$

FFDs as Linguistic summaries

1999: Ben Yahia, Jaoua:
FFD or The linguistic summary

A FFD, denoted by $(A, L_i) \rightarrow_\beta (B, L_j)$, where $A, B \in R$, $\beta, \theta_t \in [0, 1]$ holds if for all $r \in \mathcal{D}$:

$$(\mu_{L_i}(r(A)) \rightarrow_L \mu_{L_j}(r(B)) \geq \theta_T,$$

where

$$\beta = \mathsf{min}(1, 1 - \mu_{L_i}(r(A)) + \mu_{L_j}(r(B)))$$

.

FFDs as constraints based on fuzzy rules

1992: Dubois, Prade

- Variables x, y on U, V representing by possibility distributions
- A, B fuzzy sets on U, V
- Certainty rule:
  A is certain for x: for all $u \in U : \pi_x(u) \leq \mu_A(u)$.
  "The more x is A, the more certain y lies in B".

$$\forall u \in U, v \in V : \pi_{y|x}(v, u) \leq \mathsf{max}(\mu_B(v), 1 - \mu_A(u))$$

Certainty rule - type dependency
Let u, u', v, v' be values of $x = r_1(A), x' = r_2(A)$, $y = r_1(B), y' = r_2(B)$, R, S similarity relations
"The more similar (in the sense of R) $r_1(A)$ and $r_2(A)$ are, the more *certain* the similarity (as described by S) of $r_1(B)$ and $r_2(B)$ is."

$$\Pi_{y,y'|x,x'}(v, v', u, u') \leq \mathsf{max}(v \approx_S v', 1 - u \approx_R u')$$

FFDs as constraints based on fuzzy rules

- Variables x, y on U, V representing by possibility distributions
- A, B fuzzy sets on U, V
- Possibility rule: B is possible range for y: for all $v \in V : \mu_B(v) \leq \pi_y(v)$
  "The more x is A, the more possible B is a range for y."

$$\forall u \in U, v \in V : \min(\mu_A(u), \mu_B(v)) \leq \pi_{y|x}(v, u)$$

Possibility rule - type dependency
Let u, u', v, v' be values of $x = r_1(A), x' = r_2(A)$, $y = r_1(B), y' = r_2(B)$, R, S similarity relations
"The more similar (in the sence of R) $r_1(A)$ and $r_2(A)$ are, the more *possible* the similarity (as described by S) of $r_1(B)$ and $r_2(B)$ is."

$$\Pi_{y,y'|x,x'}(v, v', u, u') \geq \min(\mu_S(v, v'), \mu_R(u, u'))$$

Conclusions

1. Many papers, many approaches, different quality
2. What they want to capture: "The closer the A values, the closer the B values."
3. Focused on FD separately
4. FD is true or not true
5. Crisp semantic entailment
6. Meaning of the rank
7. Additional parameters

Thank you