# P. Cordero, M. Enciso, L. Urbanova: Similarity-based FD - a comparative survey

Department of Computer Science, Palacky University, Olomouc (17. listopadu 12, CZ–77146 Olomouc, Czech Republic)

## Introduction

E.F. Codd: A relational model of data for large shared data banks, Communications of the ACM
- based on predicate logic and set theory

Functional dependency $A \Rightarrow B$:
- "For every pair of tuples: If they agree on attributes from $A$, then they also agree on the attributes from $B$."
- Relation to function

Many different extensions of the relational model $\Rightarrow$ many definitions of similarity-based FD (FD over domains with similarities) $\Rightarrow$ are they "proper" generalizations? What does it mean?

- Similarity-based functional dependencies: SBFD

## Extensions of Codd's model:

Similarity-based approaches (from equality to similarity)

:: domains are additionally equipped with similarity (reflexive and symmetric relation)

:: the degree of similarity:
- $[0,1]$ - most approaches: Buckles, Bosc, Cubero, Chen, Dubois, Prade, Raju, Majumdar, Rasmussen, Yager,...
- commutative semiring - Hajdinjak
- residuated lattice - Belohlavek, Vychodil; Cordero et al

Rank-based approaches (from relation to fuzzy relation)

:: data table - fuzzy set of tuples (instead of set of tuples)

:: additional column - rank, also called (membership) grade: degree to which a tuple belongs to a data table

:: rank usually takes value
- $[0,1]$ - most approaches
- commutative semiring - Green
- residuated lattice - Belohavek, Vychodil; Cordero et al
- possibility distribution on $[0,1]$ - Umano (1983)

:: meaning of the rank differs among approaches

Data extensions (from crisp to uncertain data)

- set of (possible) values - Buckles, Petry ...
- possibility distribution - Dubois, Prade, Chen, Umano, ...
- often the name fuzzy database is used

C.J.Date: " ...the domains over which relations are defined can be of arbitrary complexity ....The idea that the relational model could handle only rather simple kinds of data (like numbers and strings and dates and times) is a huge misconception, and always was. ..."

## Once you start with similarity

Classical definition of functional dependency:
$||A \Rightarrow B||_{\mathcal{D}} = \min\{(r_1(A) = r_2(A)) \to (r_1(B) = r_2(B)) \mid r_1, r_2 \in \mathcal{D}\}$.

$||A \Rightarrow B||_{\mathcal{D}} \in \{0,1\}$: degree to which FD $A \Rightarrow B$ is true in relation $\mathcal{D}$.

Equality replaced by similarity $\Rightarrow$ Several issues arise:

:: [Sim] From what set the similarities should be taken?

:: [Imp] What implication should be used? $r_1(A) \approx r_2(A)$, $r_1(B) \approx r_2(B)$ are degrees from the previously chosen set.

:: [TrFD] Should the degree to which FD is true remains crisp? Meaning should the degree come from the set $\{0,1\}$?

:: [Rank] How should the similarity-based queries like "show me all restaurants with location similar to some fix value" evaluated?

- From the logical point of view: replacing two-valued identity relations by many-valued ones (similarity) $\Rightarrow$ switching from a two-valued logic to an appropriate fuzzy logic.

Notation
- A complete residuated lattice $\mathbf{L} = \langle L, \wedge, \vee, \otimes, \to, 0, 1\rangle$
- Łukasiewicz, Gödel, Goguen implication: $\to_{Ł}$, $\to_G$, $\to_\Pi$
- Resher-Gaines implication: $\to_{RG}$

## Fuzzy function

Let $\mathcal{D} : \Pi_{y \in R} D_y \to L$ be a ranked data table and $A, B \subseteq R$. Let $\approx_A$ and $\approx_B$ be $\mathbf{L}$-similarities on corresponding domains. The degree to which $\mathcal{D}$ is a fuzzy function with respect to attributes $A$ and $B$ is defined as:

$$\mathrm{Fun}(\mathcal{D}, A, B) = \bigwedge_{r_1, r_2 \in \mathcal{D}} \Big( (\mathcal{D}(r_1) \otimes \mathcal{D}(r_2) \otimes (r_1(A) \approx_A r_2(A))) \to (r_1(B) \approx r_2(B)) \Big)$$

The following criteria will give us the degree to which: "For all relations $\mathcal{D}$: If the SBFD $A \Rightarrow B$ is satisfied by relation $\mathcal{D}$, then $\mathcal{D}$ defines a (fuzzy) function from $A$ to $B$".

$$\mathbb{F}(A \Rightarrow B, \mathrm{Fun}) = \bigwedge_{\mathcal{D}:\ \prod_{y \in R} D_y \to L} \big( ||A \Rightarrow B||_{\mathcal{D}} \to \mathrm{Fun}(\mathcal{D}, A, B) \big)$$

- the result depends on the concrete definition of SBFD
- another way of comparison

## Overview of SBFDs

- not exhaustive

(1) Buckles and Petry - 1982:
- data: sets of possible values
- (fuzzy) equivalence relation, taking values from $L = [0,1]$
- SBFD $A \Rightarrow_\beta B$ is satisfied iff for every pair of tuples $r_1, r_2$:

$$\beta * r_1(A) \approx_A r_2(A) \leq r_1(B) \approx_B r_2(B)$$

where $\beta \in [0,1]$, is called linguistic strength, * is arithmetic multiplication

(2) Prade and Testemale - 1984:
- data: possibility distribution, no ranks
- SBFD $A \Rightarrow B$ in $\mathcal{D}$ is satisfied iff for all $r_1, r_2 \in \mathcal{D}$

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

(3) Raju and Majumdar 1988:
- data: possibility distribution; ranks from $[0,1]$
- SBFD $A \Rightarrow B$ is satisfied iff for all $r_1, r_2 \in \mathcal{D}$

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

(4) Cubero et al
- Prade and Testemale model
- SBFD: $A \Rightarrow B$ is satisfied iff for all $r_1, r_2 \in \mathcal{D}$:

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

(5) Ben Yahia, Ounalli, and Jaoua 1999:
- for Raju and Majumdars model
- SBFD: $A$ determines $B$ at degree $\beta$ if for all tuples $r_1$ and $r_2$:

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

$$\beta = \min_{r_1, r_2} (r_1(A) \approx_A r_2(A) \to_L r_1(B) \approx_B r_2(B))$$

(6) Belohlavek, Vychodil 2006:
- crisp data, similarity and ranks, complete residuated lattice
- similarity ($A, B \in L^R$):

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

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

(7) Cordero et al 2009:
- crisp data, similarity and ranks, complete residuated lattice
- ranks assigned to each attribute value
- similarity: $(r_1(A) \approx_A r_2(A)) = \bigwedge_{y_i \in A} (r_1 \approx_i r_2)$,

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

- SBFD:

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

## Conclusions

| | [Imp] | [TrFD] | [Rank] | $\mathbb{F}$ |
|---|---|---|---|---|
| (1) | $\to_{RG}$ | $\{0,1\}$ | (*) | $\beta$ |
| (2) | classical | $\{0,1\}$ | No | 0 |
| (3) | $\to_{RG}$ | $\{0,1\}$ | Yes | 1 |
| (4) | classical | $\{0,1\}$ | No | $\beta \wedge (\alpha \to 0)$ |
| (5) | $\to_{Ł}$ | $\{0\} \cup [\beta,1]$ | Yes | 1 |
| (6) | residuum | $\mathbf{L}$ | Yes | 1 |
| (7) | residuum | $[0,1]$ | Yes | 1 |

(*) The relation is defined as ordinary subset or some cross product. But after executing query ranks can appear.

## Summary

:: Many papers, many approaches, different quality

:: What they want to capture: "The closer the A values, the closer the B values."

:: Validity of FD remains bivalent

:: Crisp semantic entailment (except Belohlavek, Vychodil)

:: Meaning of the rank is usually unclear

:: Ranks are usually not involved in the definition of FD - pair of tuples which belong to relation to low degree can cause a violation of FD

:: Additional parameters involved - how to chose them?

:: Focused on FD separately

## Acknowledgment

Supported by grant no. CZ.1.07/2.3.00/20.0060 of the European Social Fund in the Czech Republic.

## References

(1) M. Demirci, Fuzzy Functions and Their Applications, Journal of Mathematical Analysis and Applications, 2000
(2) R. Belohlavek, V. Vychodil, Codd's Relational Model from the Point of View of Fuzzy Logic, Journal of Logic and Computation, 2011

International Center for Information and Uncertainty
SSIU 2014: International Spring School and Workshop "Information and Uncertainty"
http://mcin.upol.cz/SSWIU-2014/