Functional dependencies over domains with similarities
A Comparative Survey I

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

Introduction

1970: E.F. Codd: A relational model of data for large shared data banks, Communications of the ACM
- based on predicate logic and set theory

Many different extensions (over 100 papers), we focused on:

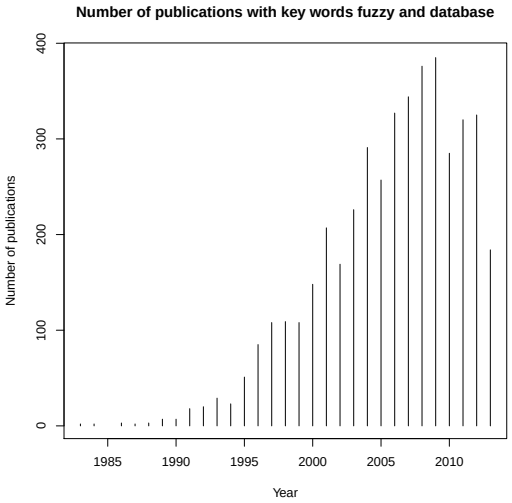
- Similarity-based approaches (from equality to similarity)
- Rank-based approaches (from relation to fuzzy relation)
- Data extensions (from crisp to uncertain data)

- we do not consider probabilistic database,...

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

Fuzzy database?

Often the names "Fuzzy database" (> 130) and "Fuzzy functional dependencies" (> 40):



Similarity-based approaches (from equality to similarity)

- domains are additionally equipped with similarity relation
- the equality relation that is presented in the original Codd's model (domain values are either "equal" or "not equal") is replaced by binary relation that maps every pair of domain values to (usually) $[0, 1]$
- the degree of similarity:
  - $[0, 1]$ - most approaches: Berzal, Bhuniya, Buckles, Bosc, Cuber, Chen, Dubois, Liu, Prade, Raju, Majumdar, Rasmussen, Yager,...
  - commutative semiring - Hajdinjak
  - De Morgan frame - Hajdinjak
  - residuated lattice - Belohlavek, Vychodil; Cordero et al
- by similarity we mean reflexive and symmetric relation (also called resemblance or proximity)

Rank-based approaches (from relation to fuzzy relation)

- data table is seen as fuzzy set of tuples (in the original model the data table is simply a set of tuples)
- data table has an additional column which contains a rank, also called (membership) grade, score or weight, to express to what degree a tuple belongs to a data table
- rank usually takes value
  - $[0, 1]$ - most approaches
  - commutative semiring - Green
  - De Morgan frame - Hajdinjak
  - residuated lattice - Belohavek, Vychodil; Cordero et al
  - possibility distribution on $[0, 1]$ - Umano (1983)
- rank is assigned to every attribute value - Medina, Cordero

Notes:
The rank is not usually involved in the definition of functional dependency.
The meaning of the rank is not clearly explained.

The meaning of the rank:

1. Compatibility with the relation: (Baldwin, Zhou)
   Rank is a "degree to which tuple satisfies the relation or is compatible with the relation". Relation Young employee (Name, Age), then a tuple belongs to the data table to degree to which it satisfy the concept "Young".
2. Global confidence level (Takahashi) :
   "(The weight is a) global confidence level in information stored in the tuple which is a part of a relation representing all or nothing concept", e.g. considering relation Likes (which is crisp!) with attributes Name and Movie, then the degree to which tuple belongs to the relation can be understood as confidence in the information stored in the tuple.
3. Degree of association between the elements of the associated tuple - Raju, Majumdar.
4. Compatibility with the set of individual constraints specified on the relation (Each employ is young.) - Mouadibb, Bonanno.
5. Estimate of the extent to which the tuple is a typical instance of the relational scheme to which it belongs.
6. Degree to which tuple matches a query - Buckles, Petry (1984); Belohlavek, Vychodil; Fagin; Rasmussen, Yager.

Data extensions (from crisp to uncertain data)

Attribute value:

- set of (possible) values - Buckles, Petry; Shenoi, Melton, Fan; Yazici; Wang,...
- possibility distribution - Bhuniya, Dubois, Prade, Chen, Umano, ...
- vague set - Zhao

Often the name Fuzzy database is used for such approaches.

C.J.Date: " ...the domains over which relations are defined can be of arbitrary complexity. As a consequence, we can have attributes of relations – or columns of tables, if you prefer – that contain geometric point, or polygons, of X rays, or XML documents, or fingerprints, or arrays, or lists, or relations, or any other kinds of values you can think of. But this idea too was always part of the relational model! 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. ..."

Codd's relational model - basic notions

1. Y– set of attributes
2. $R = \{y_1, \ldots, y_n\} \subseteq Y$– relation scheme (finite set of attributes); $A, B \subseteq R$
3. D_i – domain of attribute $y_i \in Y$ (set of all possible values)
4. $\mathcal{D} \subseteq \Pi_{y_i \in R} D_i$ – relation
5. $r_1, r_2, \ldots \in \Pi_{y_i \in R} D_i$ – tuples
6. $\approx_i$ – similarity relation on domain D_i of attribute y_i
7. $\equiv_i$ – equivalence relation on domain D_i of attribute y_i
8. binary $\mathbf{L}$–relation $\approx_i$:

(Ref) for each $u \in D_i$: $u \approx_i u = 1$,
(Sym) for each $u, v \in D_i$: $u \approx_i v = v \approx_i u$.
(Tra) for each $u, v, w \in D_i$: $u \approx_i v \otimes v \approx_i w \leq u \approx_i w$.

9. $r_1(A) \approx_A r_2(A) = \min_{y_i \in A} r_1(y_i) \approx_i r_2(y_i)$

Similarity-based approaches

1982: Buckles and Petry:

- domains are equipped with equivalence relations (called similarity in the original work) for $L = [0, 1]$ (ref, sym, $+$):

$$T1: \quad u \equiv_i w \geq \max_{v \in D_i} \{\min\{u \equiv_i v, v \equiv_i w\}\}$$

$$T2: \quad u \equiv_i w \geq \max_{v \in D_i} \{u \equiv_i v * v \equiv_i w\},$$

- tuple values are allowed to be ordinary subsets of domain with empty set excluded (not forbidden in the original model)
- relation $\mathcal{D}$ is understood as: $\mathcal{D} \subseteq \mathbf{2}^{D_1} \times \mathbf{2}^{D_2} \times \cdots \times \mathbf{2}^{D_n}$

1986: FFD $A \Rightarrow_\beta B$ holds in the Buckles/Petry model, $0 < \beta \leq 1$ iff for every pair of tuples $r_i = (d_{i1}, \ldots, d_{in}), r_j = (d_{j1}, \ldots d_{jn})$

$$\min_{y_k \in A} \{ \min_{u \in d_{ik}, v \in d_{jk}} u \equiv_k v\} \leq \beta \min_{y_r \in B} \{ \min_{u \in d_{ir}, v \in d_{jr}} u \equiv_r v\},$$

where $r_i(y_k) = d_{ik}$ is the value of attribute y_k for tuple r_i.

Similarity-based approaches II

1993: Yazici, Gocmen, Buckles, Petry - reformulation using conformance:

$$C(y_k[r_1, r_2]) = \min_{u,v \in d_{1k} \cup d_{2k}} u \equiv_k v$$

$$C(A[r_1, r_2]) = \min_{y_k \in A} C(y_k[r_1, r_2]) \quad \text{for } A \subseteq R$$

FD $A \Rightarrow B$ is satisfied iff $\forall r_1, r_2$:

$$\beta * C(A[r_1, r_2]) \leq C(B[r_1, r_2]),$$

where $\beta \in [0, 1]$ is called linguistic strength and is optional, the default value of β is 1.

Example

Assume $R = \{y_1, y_2, y_3\}$ with $D_1 = D_2 = D_3 = \{a, b, c, d\}$

$\mathcal{D}$	y_1	y_2	y_3
r_1	$\{a, b\}$	$\{c, d\}$	$\{c\}$
r_2	$\{a, b\}$	$\{c, d\}$	$\{d\}$

$||\{y_1\} \Rightarrow \{y_2\}||_{\mathcal{D}} = 1$ for $\beta = 1$? If $a \equiv_1 b > c \equiv_2 d$, then some $\beta < 1$ has to be chosen.

Possibility fuzzy data model

1983: Umano

- attribute value: possibility distribution (ambiguity in data values)
  $r_1(\text{Age}) = \{1/30, 0.8/25\}$
- rank - ambiguity in an association between values
- rank - possibility distributions on $[0, 1]$
- no similarity relations

$$\mathcal{D} = P(y_1) \times \ldots \times P(y_n) \to P([0, 1]),$$

where $P(y_i)$ is called domain and it is a collection of all possibility distributions on a universe y_i, called basic set.

Similarity-based approaches III

1984: Prade and Testemale

- possibility fuzzy data model, i.e. attribute value can be possibility distribution on domain
- every domain is extended by element e - "not applicable"
- ranks are not employed

The (fuzzy) functional dependency $A \Rightarrow B$ in $\mathcal{D}$ is satisfied iff for all $r_1, r_2 \in \mathcal{D}$

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

The FD should capture the following:
"If the values of the attribute A are equal for r_1 and r_2, we may want to express that the values of the attribute B for r_1 and r_2 cannot be far from each other".

Similarity-based approaches IV

1988: Raju and Majumdar

- relation is understood as $\mathcal{D} : \prod_{y_i \in R} D_i \to [0, 1]$
- 2 categories of imprecise data:
  Type-1: domain may be a classical set or a fuzzy set (tuple values are crisp)
  Type-2: domain may be set of fuzzy subsets (tuple value is fuzzy set)

The (fuzzy) functional dependency $A \Rightarrow B$ in $\mathcal{D}$ is satisfied iff for all $r_1, r_2 \in \mathcal{D}$

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

Using Resher-Gaines implication, $a \to_{RG} b = 1$ iff $a \leq b$, 0 otherwise.

$$r_1(A) \approx_A r_2(A) \to_{RG} r_1(B) \approx_B r_2(B)$$

- rank is not involved in the definition of FD.
- completeness of inference axioms is conditioned by the following: For each domain D_i: $\exists u, v \in D_i : u \approx_i v = 0$.

Similarity-based approaches V

1991: Chen

- possibility distribution
- similarity relation on each domain
- no ranks

The (fuzzy) FD $A \Rightarrow B$ holds in $\mathcal{D}$ iff

$$\min_{r_1, r_2 \in \mathcal{D}} (r_1(A) \approx_A r_2(A) \rightarrow_G r_1(B) \approx_B r_2(B)) \geq \theta.$$

The FD expresses the fact that: "Close B values correspond to close A values".
Later Chen et al (1992):
"Close B values correspond to close A values, and identical B values correspond to identical A values". The $\rightarrow$ is classical implication when $r_1(A)$ and $r_2(A)$ are identical, and Gödel implication otherwise.

$$\begin{aligned} \text{If} \quad & r_1(A) = r_2(A) \rightarrow r_1(B) = r_2(B) \\ \text{Else} \quad & (r_1(A) \approx_A r_2(A) \rightarrow_G r_1(B) \approx_B r_2(B)) \geq \theta, \end{aligned}$$

The soundness and completeness of Armstrong-like inference rules have been proved.

Similarity-based approaches VI

1993: Bhuniya and Niyogi -
$A \Rightarrow B$ holds in a Raju and Majumdar's model iff $\forall r_1, r_2 \in \mathcal{D}$:

$$r_1(A) \approx_A r_2(A) \leq r_1(B) \approx_B r_2(B) \quad \text{or}$$
$$(r_1(A) \approx_A r_2(A) - r_1(B) \approx_B r_2(B)) \leq 1 - \beta,$$

and $r_1(A) \approx_A r_2(A) \geq \alpha$, $r_1(B) \approx_B r_2(B) \geq \alpha$, $\alpha < \beta < 1$.
Again, note, that the rank does not influence the truthfulness of FD.

1994: Cubero et al
$A \Rightarrow B$ is satisfied iff $\forall r_1, r_2 \in \mathcal{D}$:

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

If $r_1(A)$ and $r_2(A)$ are similar at least to degree α, then $r_1(B)$ and $r_2(B)$ must be similar at least to degree β.

Similarity-based approaches VII

1999: Ben Yahia, Ounalli, and Jaoua - dynamic functional dependency
For Raju and Majumdar's model:

Dynamic FD A determines B at degree β ($A \sim\!>_\beta B$), $\beta, \theta \in [0,1]$ holds in $\mathcal{D}$ if for all tuples r_1 and r_2 we have:

$$(r_1(A) \approx_A r_2(A)) \rightarrow_L (r_1(B) \approx_B r_2(B)) \geq \theta_T$$

where

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

The parameter θ is fixed by the database designer.
Authors also proposed inference axioms and they proved the soundness. The completeness is not proved.

Similarity-based approaches VIII

1999: Bosc, Pivert, and Ughetto - classical (crisp) data, no ranks, similarity relation
They proposed two extensions to classical FD:

1. Similarity is used only in the consequence part (which is meant to express tolerance) and FD is defined as:

$$\forall r_1, r_2 \in \mathcal{D} : r_1(A) = r_2(A) \rightarrow r_1(B) \approx_B r_2(B)$$

2. Similarity relation is used in both parts

$$\forall r_1, r_2 \in \mathcal{D} : r_1(A) \approx_A r_2(A) \rightarrow r_1(B) \approx_B r_2(B)$$

   Meaning: "The closer the A values, the closer the B values."
   "Employees with similar experiences and jobs must have similar salaries."

- Residuated implication corresponding to some t-norm.
- Unfortunately, authors presented only the definitions and do not go any further.

Similarity-based approaches - continue next time

Thank you