IFCS 2006 Conference
Data Science and Classification
Knowledge Mining by Symbolic Data Analysis and the Sodas Sofware

Dissimilarity and Matching

Annalisa Appice
Donato Malerba

Department of Computer Science
University of Bari

Knowledge Acquisition & Machine Learning Lab

Why computing dissimilarity measures?

- Several data analysis techniques are based on quantifying a dissimilarity (or similarity) measure between multivariate data.
  - Visualization-based symbolic descriptions exploration
  - Classification
  - Discriminant analysis
  - Clustering
  - …
- Symbolic objects are a kind of multivariate data.
- Dissimilarity measures for both Boolean Symbolic Objects (BSOs) and Probabilistic Symbolic Objects (PSOs)
- Dissimilarity measure are computed by considering symbolic descriptions of SOs

Dissimilarity and similarity measures

- Dissimilarity Measure
  - d: $E \times E \rightarrow R$ such that
    $d^*_a = d(a,a) \leq d(a,b) = d(b,a) < \infty \quad \forall a,b \in E$
- Similarity Measure
  - s: $E \times E \rightarrow R$ such that
    $s^*_a = s(a,a) \geq s(a,b) = s(b,a) \geq 0 \quad \forall a,b \in E$

  Generally: $\forall a \in E$: $d^*_a = d^*$ and $s^*_a = s^*$ and specifically, $d^* = 0$ while $s^* = 1$
- Dissimilarity measures can be transformed into similarity measures (and vice-versa):

  $d = \phi(s) \quad (s = \phi^{-1}(d))$

  where $\phi(s)$ strictly decreasing function, and $\phi(1) = 0$, $\phi(0) = \infty$

Dissimilarity measures between BSO's

Author(s) (Year) → Notation from the ASSO Workbench

- Gowda & Diday (1991) → U_1
- Ichino & Yaguchi (1994) → U_2, U_3, U_4
- De Carvalho (1994) → SO_1, SO_2
- De Carvalho (1996, 1998) → SO_3, SO_4, SO_5, C_1
- Dissimilarity measure based on Flexible Matching → SO_6

Gowda & Diday's dissimilarity measure

U_1: $D(a, b) = \sum_{j=1}^{p} D(A_j, B_j)$

If Y_j is a continuous variable:

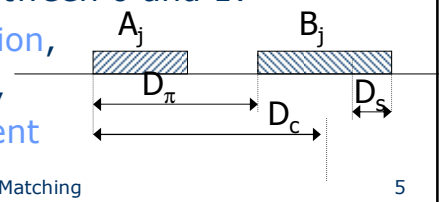
$$D(A_j, B_j) = D_\pi(A_j, B_j) + D_s(A_j, B_j) + D_c(A_j, B_j)$$

while if Y_j is a nominal variable:

$$D(A_j, B_j) = D_s(A_j, B_j) + D_c(A_j, B_j)$$

where the components are defined so that their values are normalized between 0 and 1:

- $D_\pi(A_j, B_j)$ due to position,
- $D_s(A_j, B_j)$ due to span,
- $D_c(A_j, B_j)$ due to content



Gowda & Diday's dissimilarity measure

Properties:

$D(a, b) = 0 \Rightarrow a = b$ (definiteness property),

No proof is reported for the triangle inequality property → $d(a, b) \leq d(a, c) + d(c, b)$

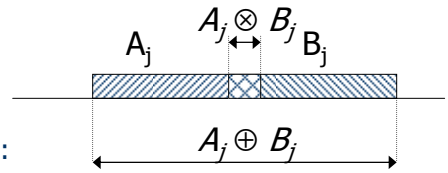
Ichino & Yaguchi's dissimilarity measures

Ichino & Yaguchi's dissimilarity measures are based on the Cartesian operators join $\oplus$ and meet $\otimes$.

For continuous variables:

$A_j \oplus B_j$

$A_j \otimes B_j$



while for nominal variables:

$A_j \oplus B_j = A_j \cup B_j$ and $A_j \otimes B_j = A_j \cap B_j$

Given a pair of subsets (A_j, B_j) of Y_j the componentwise dissimilarity $\phi(A_j,B_j)$ is:

$$\phi(A_j, B_j) = |A_j \oplus B_j| - |A_j \otimes B_j| + \gamma\,(2|A_j \otimes B_j| - |A_j| - |B_j|)$$

where $0 \le \gamma \le 0.5$ and $|A_j|$ is defined depending on variable types.

Ichino & Yaguchi's dissimilarity measures

$\phi(A_j,B_j)$ are aggregated by an aggregation function such as the generalised Minkowski's distance of order q:

U_2 $$d_q(a,b) = \sqrt[q]{\sum_{j=1}^{p} [\phi(A_j,B_j)]^q}$$

Drawback: dependence on the chosen units of measurements.

Solution: normalization of the componentwise dissimilarity:

U_3 $$d_q(a,b) = \sqrt[q]{\sum_{j=1}^{p} [\psi(A_j,B_j)]^q}, \qquad \psi(A_j,B_j) = \frac{\phi(A_j,B_j)}{|Y_j|}$$

The weighted formulation guarantees that $d_q(a,b) \in [0,1]$.

U_4 $$d_q(a,b) = \sqrt[q]{\sum_{j=1}^{p} [c_j \psi(A_j,B_j)]^q}$$

The above measures are all metrics

de Carvalho's dissimilarity measures

A straightforward extension of similarity measures for classical data matrices with nominal variables.

	Agreement	Disagreement	Total
Agreement	$\alpha=\mu(A_j \cap B_j)$	$\beta=\mu(A_j \cap c(B_j))$	$\mu(A_j)$
Disagreement	$\chi=\mu(c(A_j) \cap B_j)$	$\delta=\mu(c(A_j) \cap c(B_j))$	$\mu(c(A_j))$
Total	$\mu(B_j)$	$\mu(c(B_j))$	$\mu(Y_i)$

where $\mu(V_j)$ is either the cardinality of the set V_j (if Y_j is a nominal variable) or the length of the interval V_j (if Y_j is a continuous variable).

de Carvalho's dissimilarity measures

Five different similarity measures s_i, $i = 1, \ldots, 5$, are defined:

s_i	Comparison Function	Range	Property
s_1	$\alpha/(\alpha+\beta+\chi)$	[0,1]	metric
s_2	$2\alpha/(2\alpha+\beta+\chi)$	[0,1]	semi metric
s_3	$\alpha/(\alpha+2\beta+2\chi)$	[0,1]	metric
s_4	$\frac{1}{2}[\alpha/(\alpha+\beta)+\alpha/(\alpha+\chi)]$	[0,1]	semi metric
s_3	$\alpha/[(\alpha+\beta)(\alpha+\chi)]^{1/2}$	[0,1]	semi metric

metric → definiteness, triangle inequality

semi-metric → triangle inequality

The corresponding dissimilarities are $d_i = 1 - s_i$.

The d_i are aggregated by an aggregation function AF such as the generalised Minkowski metric, thus obtaining:

$$\text{SO\_1} \qquad d_a^i(a,b) = \sqrt[q]{\sum_{j=1}^{p} [w_j d_i(A_j, B_j)]^q} \qquad 1 \le i \le 5$$

de Carvalho's extension of Ichino & Yaguchi's dissimilarity measure

A different componentwise dissimilarity measure:

$$\psi'(A_j, B_j) = \frac{\phi(A_j, B_j)}{\mu(A_j \oplus B_j)}$$

where ϕ is defined as in Ichino & Yaguchi's dissimilarity measure.

SO_2 $$d'_q(a,b) = \sqrt[q]{\sum_{j=1}^{p} \left[\frac{1}{p} \psi'(A_j, B_j) \right]^q}$$

This measure is a metric.

The description-potential approach

All dissimilarity measures considered so far are defined by two functions:

- a comparison function (componentwise measure)
- and an aggregation function.

A different approach is based on the concept of description potential $\pi(a)$ of a symbolic object a.

$$\pi(a) = \prod_{j=1}^{p} \mu(A_j)$$

where $\mu(V_j)$ is either the cardinality of the set V_j (if Y_j is a nominal variable) or the length of the interval V_j (if Y_j is a continuous variable).

The description-potential approach

SO_3 $d'_1(a,b) = \pi(a \oplus b) - \pi(a \otimes b) + \gamma[2\pi(a \otimes b) - \pi(a) - \pi(b)]$

SO_4 $d'_2(a,b) = \dfrac{\pi(a \oplus b) - \pi(a \otimes b) + \gamma[2\pi(a \otimes b) - \pi(a) - \pi(b)]}{\pi(a^E)}$

SO_5 $d'_2(a,b) = \dfrac{\pi(a \oplus b) - \pi(a \otimes b) + \gamma[2\pi(a \otimes b) - \pi(a) - \pi(b)]}{\pi(a \oplus b)}$

The triangular inequality does not hold for SO_3 and SO_4, which are equivalent. SO_5 is a metric.

Description-potential for constrained BSO's

Given a BSO a and a logical dependence expressed by the rule:

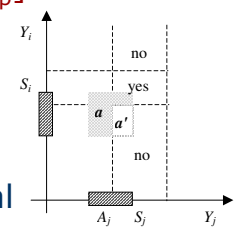
if $[Y_j = S_j]$ then $[Y_i = S_i]$

the incoherent restriction a' of a is defined as:

$a' = [Y_1=A_1] \wedge \ldots \wedge [Y_{j-1}=A_{j-1}] \wedge [Y_j=A_j \cap S_j] \wedge \ldots \wedge [Y_{i-1}=A_{i-1}] \wedge [Y_i=A_i \cap (Y_i \setminus S_i)] \wedge \ldots \wedge [Y_p=A_p]$

Then the description potential of a is:

$$\pi(a) = \prod_{j=1}^{p} \mu(A_j) - \pi(a')$$



A similar extension exists for hierarchical dependencies.

Dissimilarity measures for constrained BSO's

- The extended definition of description potential can be applied to the computation of SO_3, SO_4 and SO_5.
- de Carvalho proposed:
  - an extension of ψ', so that SO_2 can also be applied to constrained BSOs.
  - an extension of α, β, χ, and δ in order to take into account of constraints. Therefore, SO_1 can also be applied to constrained BSOs.

C_1 $$d''_q(a,b) = \sqrt[q]{\frac{\sum_{j=1}^{p} d_i(A_j,B_j)^q}{\sum_{j=1}^{p} \delta(j)}}, \text{ where } \delta(j) = \begin{cases} 0 & \text{if } Y_j = NA \\ 1 & \text{otherwise} \end{cases}$$

If all BSO's are coherent, then the dissimilarity measures do not change.

Dissimilarity measures for constrained BSO's

- The extended definition of description potential can be applied to the computation of the distances SO_3, SO_4 and SO_5.
- de Carvalho proposed an extension of ψ', so that SO_2 can also be applied to constrained BSO:

$$d'_q(a,b) = \sqrt[q]{\sum_{j=1}^{p} \left[\frac{1}{p} \psi'\_\text{constraine}\, d(A_j,B_j) \right]^q}$$

where:

$$\psi'\_\text{constrain}\, ed(A_j,B_j) = \frac{(1-\gamma)(\chi \cdot \tau + \beta \cdot \sigma) + \mu(\overline{A}_j \cap \overline{B}_j \cap (A_j \oplus B_j))}{\alpha \cdot \rho + \beta \cdot \sigma + \chi \cdot \tau + \mu(\overline{A}_j \cap \overline{B}_j \cap (A_j \oplus B_j))}$$

Dissimilarity measures for constrained BSO's

$$\rho := \frac{\bar{\pi}(a) + \bar{\pi}(b) \quad \text{under hypothesis of logical dependence between variables}}{\bar{\pi}(a) + \bar{\pi}(b) \quad \text{under hypothesis of no logical dependence between variables}}$$

where

$\bar{\bar{a}} = [Y_1=A_1] \wedge \ldots \wedge [Y_{j-1}=A_{j-1}] \wedge [Y_j=A_j \cap B_j] \wedge \ldots \wedge [Y_p=A_p]$

$\bar{\bar{b}} = [Y_1=B_1] \wedge \ldots \wedge [Y_{j-1}=B_{j-1}] \wedge [Y_j=A_j \cap B_j] \wedge \ldots \wedge [Y_p=B_p]$

Dissimilarity measures for constrained BSO's

$$\sigma := \frac{\hat{\pi}(a) \quad \text{under hypothesis of logical dependence between variables}}{\hat{\pi}(a) \quad \text{under hypothesis of no logical dependence between variables}}$$

where

$\hat{a} = [Y_1=A_1] \wedge \ldots \wedge [Y_{j-1}=A_{j-1}] \wedge [Y_j=A_j \cap c(B_j)] \wedge \ldots \wedge [Y_p=A_p]$

Dissimilarity measures for constrained BSO's

$$\tau := \frac{\hat{\pi}(b) \text{ under hypothesis of logical dependence between variables}}{\hat{\pi}(b) \text{ under hypothesis of no logical dependence between variables}}$$

where

$$\hat{b} = [Y_1=B_1] \wedge \ldots \wedge [Y_{j-1}=B_{j-1}] \wedge [Y_j=c(A_j) \cap B_j] \wedge \ldots \wedge [Y_p=B_p]$$

Dissimilarity measures for constrained BSO's

de Carvalho proposed an extension of α, β, χ in order to take into account of constraints

	Agreement	Disagreement
Agreement	$\alpha=\mu(A_j \cap B_j) \times \rho$	$\beta=\mu(A_j \cap c(B_j)) \times \sigma$
Disagreement	$\chi=\mu(c(A_j) \cap B_j) \times \tau$	

Dissimilarity measures for constrained BSO's

The previous extension of α, β, χ in order to take into account of constraints, can be used in SO_1.

C_1 $$d''_q(a,b) = \sqrt[q]{\frac{\sum_{j=1}^{p} d_i(A_j,B_j)^q}{\sum_{j=1}^{p} \delta(j)}}$$

where $$\delta(j) = \begin{cases} 0 & \text{if } Y_j = NA \\ 1 & \text{otherwise} \end{cases}$$

If all BSO's are coherent, then the dissimilarity measures do not change.

Dissimilarity measures between PSO's

Why are needed new dissimilarity measures for PSOs?

Dissimilarity measures for BSOs don't take the probabilities into account → information loss.

Dissimilarity measures for PSO are needed.

Defining dissimilarity measures for PSOs

Steps:

1. Define coefficients measuring the divergence between two probability distributions
   - Kullback-Leibler divergence (m_{KL})
   - Chi-square divergence ($m\chi$)
   - Hellinger + Cherneff' s distance of order S ($m_C^{(S)}$)
   - Renyi's divergence of order S ($m_R^{(S)}$))

   } similarity coefficient (*) } non symmetric coefficients

   - Variation distance + Minkowski's distance of order p (m_P) — symmetric coefficient

(*) from them two dissimilarity measures, namely the Renyi's and Chernoff's coefficients, are obtained

Defining dissimilarity measures for PSOs

2. Symmetrize the non symmetric coefficients

$$\underline{m}(P,Q) = m(Q,P) + m(P,Q)$$

3. Aggregate the contribution of all variables to compute the dissimilarity between two symbolic objects
   - aggregate by sum
   - aggregate by product

Defining dissimilarity measures for PSOs

Name	Componentwise dissimilarity measure	Objectwise dissimilarity measure
P_1	m_p or a symmetrized version of m_{KL}, m_χ, $m_C^{(s)}$, $m_R^{(s)}$	$\sqrt[p]{\sum_{i=1}^{m} [c_i \, m(A_i, B_i)]^p}$
P_2	$m_p(P,Q)$	$1 - \frac{\prod_{i=1}^{m} \left(\sqrt[p]{2} - \sqrt[p]{m_p(A_i,B_i)}\right)}{\left(\sqrt[p]{2}\right)^m}$

Dissimilarity measures for mixed SOs

Mixed symbolic descriptions:

1. separating the Boolean part from the Probabilistic one
2. computing dissimilarity values separately for these parts.
3. dissimilarity values obtained by comparing the Boolean parts and the Probabilistic parts respectively are then combined by sum or product.

Some applications

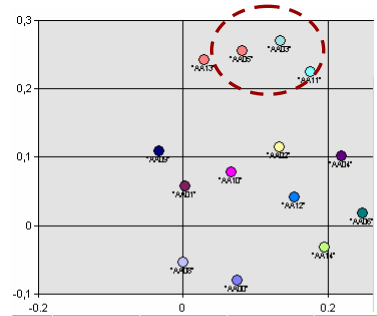
1. Visualization-based SOs exploration

- Bi-dimensional scatterplot
- Line charts

Bi-dimensional scatterplot

- Bi-dimensional from symbolic descriptions to points
- mapping based on an extension of Sammon's algorithm
  - input:
    1. n symbolic descriptions S
    2. a dissimilarity measure d
  - output:

    n 2D points whose Euclidean distances preserve the "structure" of dissimilarity matrix M on (S,d)



	"AA00"	"AA01"	"AA02"	"AA03"	
"AA00"	0.000000e+000	2.007600e-001	2.180790e-001	2.558370e-001	2.
"AA01"	2.007600e-001	0.000000e+000	1.452780e-001	2.233490e-001	1.
"AA02"	2.180790e-001	1.452780e-001	0.000000e+000	2.008580e-001	1.
"AA03"	2.558370e-001	2.233490e-001	2.008580e-001	0.000000e+000	2.
"AA04"	2.152710e-001	1.769480e-001	1.115270e-001	2.240720e-001	0.
"AA05"	2.528430e-001	2.337120e-001	2.011390e-001	4.9[illegible]e-002	2.
"AA06"	1.983090e-001	2.072110e-001	1.877840e-001	2.48[illegible]e-001	1.
"AA08"	1.045990e-001	2.069750e-001	2.373750e-001	2.655390e-001	2.

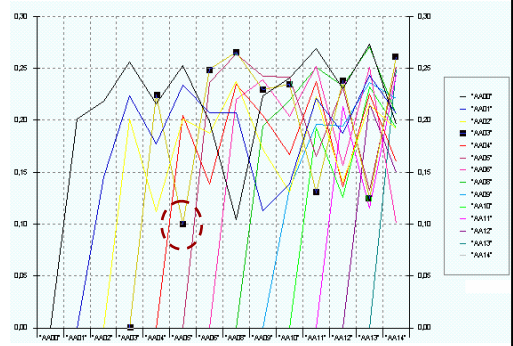
Some applications

1. Visualization-based SOs exploration

- Bi-dimensional scatterplot
- Line charts

Line-charts

- dissimilarity values are reported along the vertical axis
- individual identifiers (labels or names) are reported on the horizontal axis

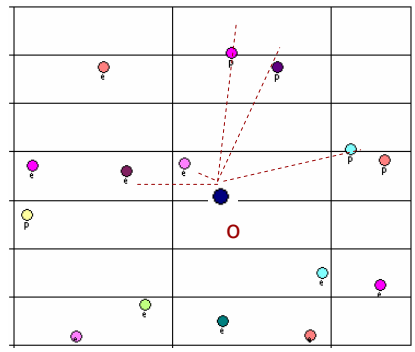


Some applications

2. Classification

- SO-NN (Symbolic Objects Nearest Neighbour)

K=5 Y={e, p} Y'(o)=0.6 p, 0.4 e



(http://www.di.uniba.it/ ~malerba/ software/SONN/index.htm)

Matching comparison

- Matching is the process of comparing two or more structures to discover their similarities or differences.
- Similarity judgments in the matching process are directional. They have
  - a referent a, i.e., a SO representing a class of objects
    r: [profession={farmer, driver}] ∧ [age=[24,34]]
  - a subject b, i.e., a SO corresponding to the description of an individual
    s: [profession=farmer] ∧ [age=28]
- Matching two structures is a common problem to many domains, like symbolic classification, pattern recognition, data mining and expert systems.
  - Matching BSOs
  - Matching PSOs

Matching BSO's

Given two BSO's a and b, the matching operators define whether the subject b is the description of an individual in the extension of the referent a.

- Two matching operators for BSO's have been defined:
  - Canonical matching
  - Flexible matching

Canonical matching operator

- The result of the canonical matching operator is either 0 (false) or 1 (true).
- If E denotes the space of BSO's described by a set of p variables Y_i taking values in the corresponding domains Y_j, then the matching operator is a function:

$$\text{Match: } E \times E \rightarrow \{0, 1\}$$

such that for any two BSO's a, b $\in$ E:

$$a = [Y_1=A_1] \wedge [Y_2=A_2] \wedge \ldots \wedge [Y_p=A_p]$$
$$b = [Y_1=B_1] \wedge [Y_2=B_2] \wedge \ldots \wedge [Y_p=B_p]$$

it happens that:

- Match(a,b) = 1 if $B_j \subseteq A_j$ for each j=1, 2, …, p
- Match(a,b) = 0 otherwise.

Canonical matching operator

Example:

District1 = [profession={farmer, driver}] ∧ [age=[24,34]]
Indiv1 = [profession=farmer] ∧ [age=28]
Indiv2 = [profession=salesman] ∧ [age=[27,28]]

Match(District1, Indiv1) = 1
Match(District1, Indiv2) = 0

Canonical matching operator

- The canonical matching function satisfies two out of three properties of a similarity measure:
  - ∀ a, b ∈ E: Match(a, b) ≥ 0
  - ∀ a, b ∈ E: Match(a, a) ≥ Match(a, b)

while it does not satisfy the *commutativity* or *simmetry* property

- ∀ a, b ∈ E: Match(a, b) = Match(b, a)

because of the different role played by *a* and *b*.

Flexible matching operator

Problem: The requirement $B_i \subseteq A_i$ for each i=1, 2, …, p, might be too strict for real-world problems.

Example:

District1 = [profession={farmer, driver}] ∧ [age=[24,34]]

Indiv3 = [profession=farmer] ∧ [age=23]

Match(District1,Indiv3) = 0

A flexible definition of matching operator that returns a number in [0,1] corresponding to the degree of match between two BSO's, that is

flexible-matching: E × E → [0,1]

Flexible matching operator

For any two BSO's a and b,

1. flexible-matching(a,b)=1 if Match(a,b)=true,
2. flexible-matching(a,b)→[0,1) otherwise.

→ probability of a matching b provided that a change is made in b.

Let $E_a = \{b' \in E | Match(a,b')=1\}$ and P(b|b') be the conditional probability of observing b given that the original observation was b'. Then

$$\text{flexible-matching}(a,b) \underset{\text{def}}{=} \max_{b' \in E_a} P(b \mid b')$$

that is flexible-matching(a,b) equals the maximum conditional probability over the space of BSO's canonically matched by a.

Flexible matching operator

Since $b(b') = b_1 \wedge \ldots \wedge b_p\ (b'_1 \wedge \ldots \wedge b'_p)$

By assuming b_i depends exclusively on b'_i then

$$P(b \mid b') = \prod_{i=1\ldots p} P(b_i \mid b'_i)$$

$b_i(b'_i)$: $[Y_i = v_i]([Y_i = v'_i])$ then

$$P(b_i \mid b'_i) = P([Y_i = v_i] \mid [Y_i = v'_i]) = P(\delta_I(v'_i, Y) \geq \delta_I(v'_i, v_i))$$

Example

$$P([Y_i = v_i] \mid [Y_i = v'_i]) = \frac{|Y_j| - 1}{|Y_j|}$$ where Y_j is nominal

Flexible matching operator

Flexible matching definition can be generalized to the case of comparing any pair of BSOs → not necessarily a BSO describing a class against a BSO describing an individual

$$\text{flexMatch}\ (a, b) = \max_{b' \in E(a)} \prod_{i=1\ldots p} \sum_{j=1\ldots q} \frac{1}{q} P(b_{ij} \mid b'_i)$$

where q is the number of categories for variable j in b

Flexible matching: an example

$a = [Y_1 \in \{yellow, green, white\}] \wedge [Y_2 \in \{Ford, Fiat, Mercedes\}]$

$b = [Y_1 \in \{yellow, black\}] \wedge [Y_2 \in \{Fiat, Audi\}]$

Y_1 = {yellow, red, green, white, black} is the domain of Y_1

Y_2 = { Ford, Fiat, Mercedes, Audi, Peugeot, Renault}

E_a={$b_1' = [Y_1 = yellow] \wedge [Y_2 = Ford]$; $b_2' = [Y_1 = yellow] \wedge [Y_2 = Fiat]$;
$b_3' = [Y_1 = yellow] \wedge [Y_2 = Mercedes]$; $b_4' = [Y_1 = green] \wedge [Y_2 = Ford]$
$b_5' = [Y_1 = green] \wedge [Y_2 = Fiat]$; $b_6' = [Y_1 = green] \wedge [Y_2 = Mercedes]$;
$b_7' = [Y_1 = white] \wedge [Y_2 = Ford]$; $b_8' = [Y_1 = white] \wedge [Y_2 = Fiat]$;
$b_9' = [Y_1 = white] \wedge [Y_2 = Mercedes]$}

$P(b_{11} \mid b_{11}') = P(Y_1 = yellow \mid Y_1 = yellow) = 1$

$P(b_{12} \mid b_{11}') = P(Y_1 = black \mid Y_1 = yellow) = 4/5$

$P(b_1 \mid b_1') = 0.5(P(b_{11} \mid b_{11}') + P(b_{12} \mid b_{11}')) = 0.9$

…

$P(b2|b1') = 0.833$

$P(b_1 \mid b_1') \times P(b2|b1') = 0.75$ → **flexMatch(a,b)>0.75**

Matching PSOs

Flexible matching operator can be extended to the case of PSOs.

1. $\text{FlexMatch}(a,b) = \max_{b' \in E'_a} \prod_{i=1\ldots p} P(b_i') \sum_{j=1..q} P(b_{ij}) P(b_{ij} | b'_i)$

2. or alternatively:

$$\text{FlexMatch}(a,b) = \prod_{i=1\ldots p} f(A_i, B_i)$$

where f is the KL-divergence, χ^2-divergence or Hellinger coefficient.

PROBLEM: KL-divergence and χ^2-divergence are dissimilarity measures → Similarity values can be obtained as:

$$f(A, B) = e^{-x}$$

where x denotes either the KL-divergence value or the χ^2-divergence

Matching mixed SOs

Mixed SOs:

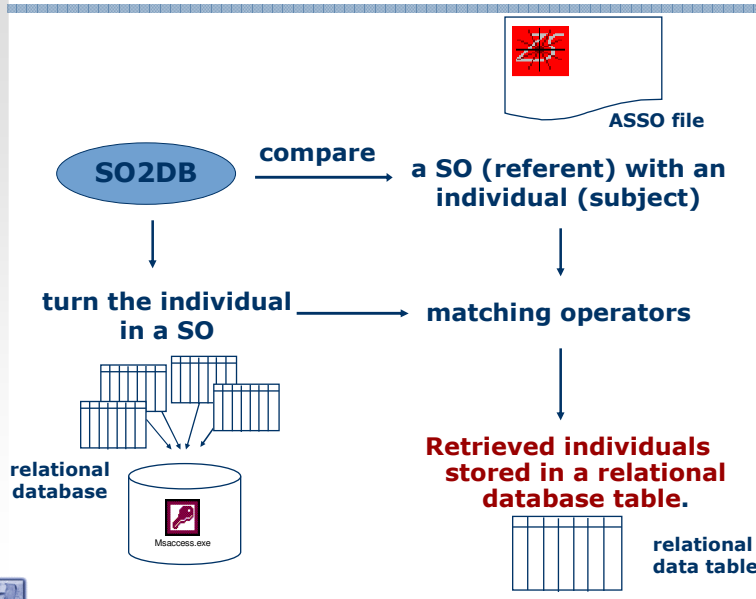
1. separating the BSOs from the PSOs
2. computing matching values separately for these SOs.
3. Matching values obtained by comparing the Boolean parts and the Probabilistic parts respectively are then combined by product.

An application

Retrieve the individuals in a relational database which correspond to some characteristics expressed by means of a SO or a set of SO's

→ SO2DB (From Symbolic Objects to DataBase)

SO2DB



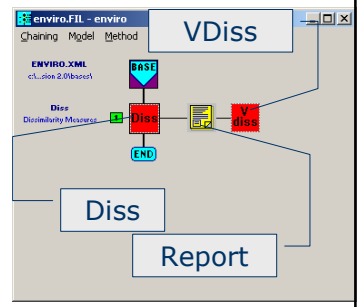
Flexible matching based dissimilarity measure

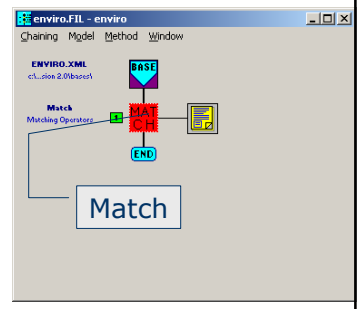
$$d(a,b) = 1 - \frac{flexMatch(a,b) + flexMatch(b,a)}{2}$$

where a, b either BSOs or PSOs.

Dissimilarity and Matching

- Both dissimilarity measures and matching operators are available in the method Dissimilarity and Matching of the ASSO software.
- Input: ASSO file
- Output for dissimilarity measures:
  - Report
  - ASSO file with dissimilarity matrix
  - VDiss: dissimilarity matrix based SOs visualization
    - Bi-dimensional scatterplot
    - Line charts
- Output for matching operators:
  - Report
  - ASSO file with matching matrix
- Developer Dipartimento di Informatica, University of Bari, Italy.





References

Esposito F., Malerba D., V. Tamma, H.H. Bock. Classical resemblance measures. Chapter 8.1

Esposito F., Malerba D., V. Tamma. Dissimilarity measures for symbolic objects. Chapter 8.3

Esposito F., Malerba D., & F.A. Lisi. Matching symbolic objects. Chapter 8.4

in H.-H. Bock, E. Diday (eds.): Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data. Springer Verlag, Heidelberg, 2000.

Malerba D., Sanarico L., & V. Tamma. A comparison of dissimilarity measures for Boolean symbolic data. In P. Brito, J. Costa, & D. Malerba (Eds.), Proc. of the ECML 2000 Workshop on Dealing with Structured Data in Machine Learning and Statistics, Barcelona, 2000.

Malerba D., Esposito F., Gioviale V., & V. Tamma. Comparing Dissimilarity Measures in Symbolic Data Analysis. Pre-Proceedings of EKT-NTTS, vol. 1, pp. 473-481.

Malerba D., Esposito D., Didonna, M. A. Gioviale, V., & S. Presta SO2DB and propagation on Data Base. Internal Report of ASSO Project, 2001.

Malerba D., Esposito F., & M. Monopoli Comparing dissimilarity measures for Probabilistic Symbolic Objects. In Data Mining III, eds. A. Zanasiet al. (eds.), Series Management Information Systems, Vol. 6, pp. 31-40. Southampton: WIT Press, 2002.

Esposito F., Malerba D., & A. Appice. Dissimilarity and Matching. Internal Report of ASSO Project, 2005.

Malerba D., Esposito F., A. Appice Symbolic Object exportation to a Database. Internal Report of ASSO Project, 2005.

Appice A. , D'Amato C. , Esposito F. & D. Malerba. Classification of Symbolic Objects: A Lazy Learning Approach. *Intelligent Data Analysis*, in press, 2006.

Thanks

…for your attention…

…Questions?