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**
  - \( d: \mathbb{E} \times \mathbb{E} \rightarrow \mathbb{R} \) such that
    \[
    d^*_a = d(a,a) \leq d(a,b) = d(b,a) < \infty \quad \forall a,b \in \mathbb{E}
    \]

- **Similarity Measure**
  - \( s: \mathbb{E} \times \mathbb{E} \rightarrow \mathbb{R} \) such that
    \[
    s^*_a = s(a,a) \geq s(a,b) = s(b,a) \geq 0 \quad \forall a,b \in \mathbb{E}
    \]
    Generally: \( \forall a \in \mathbb{E}: d^*_a = d^* \) 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) \rightarrow Notation from the ASSO Workbench**

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

\[ U_{-1}: \quad D(a, b) = \sum_{j=1}^{p} D(A_j, B_j) \]

If \( Y_j \) is a continuous variable:
\[ D(A_j, B_j) = D_p(A_j, B_j) + D_s(A_j, B_j) + D_c(A_j, B_j) \]
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:
- \( D_p(A_j, B_j) \) due to position,
- \( D_s(A_j, B_j) \) due to span,
- \( D_c(A_j, B_j) \) due to content

Properties:
\[ D(a, b) = 0 \Rightarrow a = b \] (definiteness property),
No proof is reported for the triangle inequality property \( \Rightarrow d(a, b) \leq d(a, c) + d(c, b) \)
Ichino & Yaguchi’s dissimilarity measures

Ichino & Yaguchi’s dissimilarity measures are based on the Cartesian operators join \( \oplus \) and meet \( \otimes \).

For continuous variables:
\[
\begin{align*}
A_j \oplus B_j &= A_j \\
A_j \otimes B_j &= B_j
\end{align*}
\]

while for nominal variables:
\[
\begin{align*}
A_j \oplus B_j &= A_j \cup B_j \quad \text{and} \quad A_j \otimes B_j = A_j \cap B_j
\end{align*}
\]

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 \(0 \leq \gamma \leq 0.5\) and \(|A_j|\) is defined depending on variable types.

\[
U_4 = A_j \oplus B_j
\]

\[
U_3 = d_q(a, b) = q \left( \sum_{j=1}^{p} \left[ \frac{\psi(A_j, B_j)}{|Y_j|} \right]^q \right)^{1/q}
\]

\[
U_2 = d_q(a, b) = \sqrt[2q]{q \sum_{j=1}^{p} \left[ \frac{\psi(A_j, B_j)}{|Y_j|} \right]^q}
\]

\[
\psi(A_j, B_j) = \phi(A_j, B_j)
\]

\[
\psi(A_j, B_j) = \frac{\phi(A_j, B_j)}{|Y_j|}
\]

Drawback: dependence on the chosen units of measurements.
Solution: normalization of the componentwise dissimilarity:

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

The above measures are all metrics.
de Carvalho’s dissimilarity measures

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

\[
\begin{array}{ccc}
\text{Agreement} & \text{Disagreement} & \text{Total} \\
\alpha = \mu(A_i \cap B_j) & \beta = \mu(A_i \cap c(B_j)) & \mu(A_i) \\
\gamma = \mu(c(A_i) \cap B_j) & \delta = \mu(c(A_i) \cap c(B_j)) & \mu(c(A_i)) \\
\mu(B_j) & \mu(c(B_j)) & \mu(Y_i)
\end{array}
\]

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 \(s_i, i = 1, ..., 5\), are defined:

\[
\begin{align*}
s_1 & = \frac{\alpha}{\alpha + \beta + \chi} & \text{Range} & = & [0,1] & \text{Property} & = & \text{metric} \\
s_2 & = \frac{2\alpha}{2\alpha + \beta + \chi} & [0,1] & \text{semi metric} \\
s_3 & = \frac{\alpha}{\alpha + 2\beta + 2\chi} & [0,1] & \text{metric} \\
s_4 & = \frac{1}{2}[\alpha/(\alpha + \beta) + \alpha/(\alpha + \chi)] & [0,1] & \text{semi metric} \\
s_5 & = \frac{\alpha}{[(\alpha + \beta)(\alpha + \chi)]^{\frac{1}{2}}} & [0,1] & \text{semi metric}
\end{align*}
\]

metric \(\rightarrow\) definiteness, triangle inequality

semi-metric \(\rightarrow\) 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:

\[
\text{SO}_1 \quad d^i_a(a, b) = \sqrt[1}{\sum_{j=1}^{p} w_j d(A_j, B_j)}^{1} \quad 1 \leq i \leq 5
\]
**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 \( \phi \) is defined as in Ichino & Yaguchi’s dissimilarity measure.

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

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  \( d'_i(a,b) = n(a \oplus b) - n(a \otimes b) + \gamma [2n(a \otimes b) - n(a) - n(b)] \)

SO_4  \( d'_i(a,b) = \frac{n(a \otimes b) - n(a \otimes b) + \gamma [2n(a \otimes b) - n(a) - n(b)]}{n(a^e)} \)

SO_5  \( d'_i(a,b) = \frac{n(a \otimes b) - n(a \otimes b) + \gamma [2n(a \otimes b) - n(a) - n(b)]}{n(a \otimes b)} \)

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:

\[ \text{if } [Y_j = S_j] \text{ then } [Y_i = S_i] \]

the incoherent restriction \( a' \) of \( a \) is defined as:

\[ a' = [Y_1 = A_j] \land ... \land [Y_{i-1} = A_{i-1}] \land [Y_i = A_j, S_j] \land ... \land [Y_{i-1} = A_{i-1}] \land [Y_i = A_i \cap (Y_i \setminus S_j)] \land ... \land [Y_p = A_p] \]

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 $\psi'$, so that SO_2 can also be applied to constrained BSOs.
  - an extension of $\alpha$, $\beta$, $\chi$, and $\delta$ in order to take into account of constraints. Therefore, SO_1 can also be applied to constrained BSOs.

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

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 $\psi'$, so that SO_2 can also be applied to constrained BSO:

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

where:

\[
\psi'_\text{constraine} d(A_j, B_j) = \frac{(1 - \gamma)(\chi \cdot \tau + \beta \cdot \sigma) + \mu(\overline{A_j} \cap B_j \cap (A_j \oplus B_j))}{\alpha \cdot \rho + \beta \cdot \sigma + \chi \cdot \tau + \mu(\overline{A_j} \cap B_j \cap (A_j \oplus B_j))}
\]

Dissimilarity and Matching
Dissimilarity measures for constrained BSO’s

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

where
\[ \bar{a} = [Y_1 = A_1] \land \ldots \land [Y_{j-1} = A_{j-1}] \land [Y_j = A_j \cap B_j] \land \ldots \land [Y_p = A_p] \]
\[ \bar{b} = [Y_1 = B_1] \land \ldots \land [Y_{j-1} = B_{j-1}] \land [Y_j = A_j \cap B_j] \land \ldots \land [Y_p = B_p] \]

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

where
\[ \hat{a} = [Y_1 = A_1] \land \ldots \land [Y_{j-1} = A_{j-1}] \land [Y_j = A_j \cap c(B_j)] \land \ldots \land [Y_p = A_p] \]
Dissimilarity measures for constrained BSO’s

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

where

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

---

Dissimilarity measures for constrained BSO’s

de Carvalho proposed an extension of \( \alpha, \beta, \chi \) in order to take into account of constraints

\[
\begin{align*}
\text{Agreement} & : \alpha = \mu(A_j \cap B_j) \times \rho \\
\text{Disagreement} & : \beta = \mu(A_j \cap c(B_j)) \times \sigma \\
\text{Agreement} & : \chi = \mu(c(A_j) \cap B_j) \times \tau
\end{align*}
\]
Dissimilarity measures for constrained BSO’s

The previous extension of $\alpha$, $\beta$, $\chi$ in order to take into account of constraints, can be used in SO_1.

$$C_1 d_c(a,b) = \sqrt{\frac{\sum_{j=1}^{q} d(A_j, B_j)^2}{\sum_{j=1}^{q} \delta(j)}}$$

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 $\rightarrow$ 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 + Chernoff's distance of order S (m_{c(S)})
   - Renyi's divergence of order S (m_{R(S)})
   - Variation distance + Minkowski's distance of order p (m_p)

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

2. Symmetrize the non symmetric coefficients
   \[ m(P,Q) = m(Q,P) + m(P,Q) \]

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Componentwise dissimilarity measure</th>
<th>Objectwise dissimilarity measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_1</td>
<td>m_p or a symmetrized version of m_{KL}, m_{KL'}, m_{C^s}, m_{R^s}</td>
<td>\sqrt[p]{\sum_{i=1}^{m} [c_i - m(A_i, B_i)]^p}</td>
</tr>
<tr>
<td>P_2</td>
<td>m_p(P,Q)</td>
<td>1 - \prod_{i=1}^{m} \left(1-\frac{m_p(A_i, B_i)}{\left(\frac{\kappa}{2}\right)^m}\right)</td>
</tr>
</tbody>
</table>

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)

Line-charts
   - dissimilarity values are reported along the vertical axis
   - individual identifiers (labels or names) are reported on the horizontal axis
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
    \( r: [\text{profession} = \{\text{farmer, driver}\}] \land [\text{age}=24,34] \)
  - a subject \( b \), i.e., a SO corresponding to the description of an individual
    \( s: [\text{profession}=\text{farmer}] \land [\text{age}=28] \)

- 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 $Y_j$ 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 $a, b \in E$:

\[
\begin{align*}
  a &= [Y_1=A_1] \land [Y_2=A_2] \land \ldots \land [Y_p=A_p] \\
  b &= [Y_1=B_1] \land [Y_2=B_2] \land \ldots \land [Y_p=B_p]
\end{align*}
\]

it happens that:

\[
\begin{align*}
  \text{Match}(a,b) &= 1 & \text{if } B_j \subseteq A_j & \text{for each } j=1, 2, \ldots, p \\
  \text{Match}(a,b) &= 0 & \text{otherwise.}
\end{align*}
\]
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 symmetry 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 $i = 1, 2, \ldots, p$, might be too strict for real-world problems.

**Example:**

- District1 = [profession={farmer, driver}] $\land$ [age=[24,34]]
- Indiv3 = [profession=farmer] $\land$ [age=23]

$$\text{Match(District1, Indiv3)} = 0$$

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

$$\text{flexible-matching: } E \times E \rightarrow [0,1]$$

---

**Flexible matching operator**

For any two BSO’s a and b,

1. $\text{flexible-matching}(a,b)=1$ if $\text{Match}(a,b)=\text{true}$,
2. $\text{flexible-matching}(a,b)\rightarrow[0,1)$ otherwise.

$\rightarrow$ probability of a matching $b$ provided that a change is made in $b$.

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

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

that is $\text{flexible-matching}(a,b)$ equals the maximum conditional probability over the space of BSO’s canonically matched by $a$. 

Flexible matching operator

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

By assuming \( b_i \) depends exclusively on \( b'_i \) then

\[
P(b \mid b') = \prod_{i=1}^{p} P(b_i \mid b'_i)
\]

\( b_i(b'_i): [Y_i=v_i], [Y_i=v'_i] \) 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

Dissimilarity and Matching 37

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

\[
\text{flexMatch} \ (a, b) = \max_{b' \in E(a)} \prod_{i=1}^{p} \sum_{j=1}^{q} \frac{1}{q} P(b_{ij} \mid b'_{i})
\]

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

Dissimilarity and Matching 38
Flexible matching: an example

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

\[ b = [Y_1 \in \{\text{yellow, black}\}] \land [Y_2 \in \{\text{Fiat, Audi}\}] \]

\[ Y_1 = \{\text{yellow, red, green, white, black}\} \] is the domain of \( Y_1 \)

\[ Y_2 = \{\text{Ford, Fiat, Mercedes, Audi, Peugeot, Renault}\} \]

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

\[ P(b_{11} \mid b_{11}') = P(Y_1 = \text{yellow} \mid Y_1 = \text{yellow}) = 1 \]

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

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

\[ P(b_2 \mid b_1') = 0.833 \]

\[ P(b_1 \mid b_1') \times P(b_2 \mid b_1') = 0.75 \rightarrow \text{flexMatch}(a, b) > 0.75 \]

Matching PSOs

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

1. \( \text{FlexMatch}(a, b) = \max_{b' \in E_a} \prod_{i=1}^{p} P(b_i') \sum_{j=1}^{q} P(b_j) P(b_j \mid b_i') \)

2. or alternatively:
\[
\text{FlexMatch}(a, b) = \prod_{i=1}^{p} f(A_i, B_i)
\]

where \( f \) is the KL-divergence, \( \chi^2 \)-divergence or Hellinger coefficient.

**PROBLEM:** KL-divergence and \( \chi^2 \)-divergence are dissimilarity measures \( \rightarrow \) Similarity values can be obtained as:

\[ f(A, B) = e^{-x} \]

where \( x \) denotes either the KL-divergence value or the \( \chi^2 \)-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

- SO2DB compare a SO (referent) with an individual (subject)
- turn the individual in a SO
- matching operators
- Retrieved individuals stored in a relational database table.

Flexible matching based dissimilarity measure

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

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. 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
Thanks

...for your attention...

...Questions?