# Lösung - Übungsblatt 5

## (Duplikaterkennung)

Fabian Panse

panse@informatik.uni-hamburg.de

Universität Hamburg

Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

# Aufgabe 1: Verifikation

Gegeben: $\mathcal{C}_{\text{gold}} = \{\langle 1,5,7\rangle, \langle 2,4\rangle, \langle 8,10\rangle, \langle 3,9\rangle, \langle 6\rangle\}$

$\mathcal{C} = \{\langle 1,5\rangle, \langle 2,4,7\rangle, \langle 6,8\rangle, \langle 3\rangle, \langle 9\rangle, \langle 10\rangle\}$

$M(\mathcal{C}_{\text{gold}}) = \{\{1,5\},\{1,7\},\{5,7\},\{2,4\},\{8,10\},\{3,9\}\}$

$M(\mathcal{C}) = \{\{1,5\},\{2,7\},\{4,7\},\{2,4\},\{6,8\}\}$

$TP(\mathcal{C}) = M(\mathcal{C}_{\text{gold}}) \cap M(\mathcal{C}) = \{\{1,5\},\{2,4\}\}$

$FP(\mathcal{C}) = M(\mathcal{C}) - M(\mathcal{C}_{\text{gold}}) = \{\{2,7\},\{4,7\},\{6,8\}\}$

$FN(\mathcal{C}) = M(\mathcal{C}_{\text{gold}}) - M(\mathcal{C}) = \{\{1,7\},\{5,7\},\{8,10\},\{3,9\}\}$

$Recall(\mathcal{C}) = 2/(2+4) = 0.33$ $\quad$ $Precision(\mathcal{C}) = 2/(2+3) = 0.4$

$F_1\text{-}score(\mathcal{C}) = (2 \times 0.33 \times 0.4)/(0.33 + 0.4) = 0.3616$

# Aufgabe 2: Sorted Neighborhood Method

Eingabetabelle:

| ID | Vorname | Nachname | Jahr |
|---|---|---|---|
| 1 | Florian | Panse | 1983 |
| 2 | Marko | Meier | 1989 |
| 3 | Fabian | Pause | 1981 |
| 4 | Felix | Mueller | 1985 |
| 5 | Mirko | Maier | 1990 |
| 6 | Stefan | Meier | 1979 |
| 7 | Fabian | Hanse | 1980 |
| 8 | Norbert | Ritter | 1968 |

# Aufgabe 2: Sorted Neighborhood Method

Schlüsselerzeugung:

| ID | Vorname | Nachname | Jahr | Schlüssel |
|---|---|---|---|---|
| 1 | Florian | Panse | 1983 | flpa83 |
| 2 | Marko | Meier | 1989 | mame89 |
| 3 | Fabian | Pause | 1981 | fapa81 |
| 4 | Felix | Mueller | 1985 | femu85 |
| 5 | Mirko | Maier | 1990 | mima90 |
| 6 | Stefan | Meier | 1979 | stme79 |
| 7 | Fabian | Hanse | 1980 | faha80 |
| 8 | Norbert | Ritter | 1968 | nori68 |

# Aufgabe 2: Sorted Neighborhood Method

Sortierung:

| | | |
|---|---|---|
| 7 | faha80 | |
| 3 | fapa81 | |
| 4 | femu85 | |
| 1 | flpa83 | |
| 2 | mame89 | |
| 5 | mima90 | |
| 8 | nori68 | |
| 6 | stme79 | |

# Aufgabe 2: Sorted Neighborhood Method

Windowing:

| | | |
|---|---|---|
| 7 | faha80 | 3,4 |
| 3 | fapa81 | 4,1 |
| 4 | femu85 | 1,2 |
| 1 | flpa83 | 2,5 |
| 2 | mame89 | 5,8 |
| 5 | mima90 | 8,6 |
| 8 | nori68 | 6 |
| 6 | stme79 | - |

# Aufgabe 2: Sorted Neighborhood Method

Finaler Suchraum:

$$\{\{3,7\}, \{4,7\}, \{3,4\}, \{1,3\}, \{1,4\}, \{2,4\}, \{1,2\}, \{1,5\}, \{2,5\}, \{2,8\}, \{5,8\}, \{5,6\}, \{6,8\}\}$$

Reduktion:

Vorher: $(7 \times 8)/2 = 28$ Paare

Nachher: 13 Paare

$\Rightarrow 1 - 13/28 = 0.5357 \Rightarrow 53,57\%$ Reduktion

# Aufgabe 3: Distanzbasierte Entscheidungsmodelle

$$\vec{c}_1 = \langle 0.4, 0.7, 0.1 \rangle$$
$$\vec{c}_2 = \langle 0.5, 0.5, 0.5 \rangle$$
$$\vec{c}_3 = \langle 0.1, 0.2, 0.9 \rangle$$

Euklidische Distanz: $dst(x, y) = \sqrt{\sum_{i=1}^{3}(x_i - y_i)^2}$

$$dst(\mathbf{0}, \vec{c}_1) = \sqrt{0.16 + 0.49 + 0.01} = \sqrt{0.66} = 0.812$$
$$dst(\mathbf{0}, \vec{c}_2) = \sqrt{0.25 + 0.25 + 0.25} = \sqrt{0.75} = 0.866$$
$$dst(\mathbf{0}, \vec{c}_3) = \sqrt{0.01 + 0.04 + 0.81} = \sqrt{0.86} = 0.927$$

# Aufgabe 4: Regelbasierte Entscheidungsmodelle

| **Vergleichsvektor** | **Matchingklasse** |
|---|---|
| $\vec{c}_1 = \langle 0.5, 0.2, 1.0 \rangle$ | Match |
| $\vec{c}_2 = \langle 0.9, 0.8, 0.3 \rangle$ | Match |
| $\vec{c}_3 = \langle 0.5, 0.5, 0.8 \rangle$ | Unmatch |
| $\vec{c}_4 = \langle 0.9, 0.9, 0.4 \rangle$ | Match |
| $\vec{c}_5 = \langle 0.5, 0.9, 0.4 \rangle$ | Unmatch |
| $\vec{c}_6 = \langle 0.5, 0.3, 0.7 \rangle$ | Unmatch |
| $\vec{c}_7 = \langle 0.8, 0.8, 0.9 \rangle$ | Match |
| $\vec{c}_8 = \langle 0.4, 0.6, 0.7 \rangle$ | Match |
| $\vec{c}_9 = \langle 0.7, 0.8, 0.7 \rangle$ | Match |
| $\vec{c}_{10} = \langle 0.6, 1.0, 0.2 \rangle$ | Match |
| $\vec{c}_{11} = \langle 0.6, 0.9, 0.1 \rangle$ | Match |
| $\vec{c}_{12} = \langle 0.7, 0.8, 0.5 \rangle$ | Unmatch |

# Aufgabe 4: Regelbasierte Entscheidungsmodelle

| | | |
|---|---|---|
| R1: | $c(A_1) \geq 0.8 \wedge c(A_2) \geq 0.7$ | $\Rightarrow$ Match |
| R2: | $c(A_3) \geq 0.9$ | $\Rightarrow$ Match |
| R3: | $c(A_2) \geq 0.6 \wedge c(A_3) \geq 0.7$ | $\Rightarrow$ Match |
| R4: | $c(A_1) \leq 0.5$ | $\Rightarrow$ Unmatch |
| R5: | $c(A_2) \geq 0.9$ | $\Rightarrow$ Match |
| Rdefault: | | $\Rightarrow$ Unmatch |

# Aufgabe 4: Regelbasierte Entscheidungsmodelle

| **Vergleichsvektor** | **Matchingklasse** | **Getriggerte Regeln** |
|---|---|---|
| $\vec{c}_1 = \langle 0.5, 0.2, 1.0 \rangle$ | Match | R2 + R4 |
| $\vec{c}_2 = \langle 0.9, 0.8, 0.3 \rangle$ | Match | R1 |
| $\vec{c}_3 = \langle 0.5, 0.5, 0.8 \rangle$ | Unmatch | R4 |
| $\vec{c}_4 = \langle 0.9, 0.9, 0.4 \rangle$ | Match | R1 |
| $\vec{c}_5 = \langle 0.5, 0.9, 0.4 \rangle$ | Unmatch | R4 |
| $\vec{c}_6 = \langle 0.5, 0.3, 0.7 \rangle$ | Unmatch | R4 |
| $\vec{c}_7 = \langle 0.8, 0.8, 0.9 \rangle$ | Match | R1 |
| $\vec{c}_8 = \langle 0.4, 0.6, 0.7 \rangle$ | Match | R3 + R4 |
| $\vec{c}_9 = \langle 0.7, 0.8, 0.7 \rangle$ | Match | R3 |
| $\vec{c}_{10} = \langle 0.6, 1.0, 0.2 \rangle$ | Match | R5 |
| $\vec{c}_{11} = \langle 0.6, 0.9, 0.1 \rangle$ | Match | R5 |
| $\vec{c}_{12} = \langle 0.7, 0.8, 0.5 \rangle$ | Unmatch | Rdefault |

# Aufgabe 4: Regelbasierte Entscheidungsmodelle

| **Vergleichsvektor** | **Matchingklasse** | **Getriggerte Regeln** |
|---|---|---|
| $\vec{c}_1 = \langle 0.5, 0.2, 1.0 \rangle$ | Match | R2 + ~~R4~~ |
| $\vec{c}_2 = \langle 0.9, 0.8, 0.3 \rangle$ | Match | R1 |
| $\vec{c}_3 = \langle 0.5, 0.5, 0.8 \rangle$ | Unmatch | R4 |
| $\vec{c}_4 = \langle 0.9, 0.9, 0.4 \rangle$ | Match | R1 |
| $\vec{c}_5 = \langle 0.5, 0.9, 0.4 \rangle$ | Unmatch | R4 |
| $\vec{c}_6 = \langle 0.5, 0.3, 0.7 \rangle$ | Unmatch | R4 |
| $\vec{c}_7 = \langle 0.8, 0.8, 0.9 \rangle$ | Match | R1 |
| $\vec{c}_8 = \langle 0.4, 0.6, 0.7 \rangle$ | Match | R3 + ~~R4~~ |
| $\vec{c}_9 = \langle 0.7, 0.8, 0.7 \rangle$ | Match | R3 |
| $\vec{c}_{10} = \langle 0.6, 1.0, 0.2 \rangle$ | Match | R5 |
| $\vec{c}_{11} = \langle 0.6, 0.9, 0.1 \rangle$ | Match | R5 |
| $\vec{c}_{12} = \langle 0.7, 0.8, 0.5 \rangle$ | Unmatch | Rdefault |

# Aufgabe 5: Statistische Entscheidungsmodelle

Satz 1: 100% aller Matches und 10% aller Unmatches haben den Wert 1 im ersten Attribut.

$\Rightarrow Pr(\vec{p}[1] = 1 \mid M) = 1.0$ und $Pr(\vec{p}[1] = 0 \mid M) = 0.0$

$\Rightarrow Pr(\vec{p}[1] = 1 \mid U) = 0.1$ und $Pr(\vec{p}[1] = 0 \mid U) = 0.9$

Satz 2: 40% aller Matches und 50% aller Unmatches haben den Wert 1 im zweiten Attribut.

$\Rightarrow Pr(\vec{p}[2] = 1 \mid M) = 0.4$ und $Pr(\vec{p}[2] = 0 \mid M) = 0.6$

$\Rightarrow Pr(\vec{p}[2] = 1 \mid U) = 0.5$ und $Pr(\vec{p}[2] = 0 \mid U) = 0.5$

Satz 3: 70% aller Matches und 2% aller Unmatches haben den Wert 1 im dritten Attribut.

$\Rightarrow Pr(\vec{p}[3] = 1 \mid M) = 0.7$ und $Pr(\vec{p}[3] = 0 \mid M) = 0.3$

$\Rightarrow Pr(\vec{p}[3] = 1 \mid U) = 0.02$ und $Pr(\vec{p}[3] = 0 \mid U) = 0.98$

# Aufgabe 5: Statistische Entscheidungsmodelle

Gegeben: Pattern $\vec{p}_1 = \langle 1, 0, 1 \rangle$

$$\Rightarrow Pr(\vec{p}_1 \mid M) = Pr(\vec{p}[1]{=}1 \mid M) \times Pr(\vec{p}[2]{=}0 \mid M) \times Pr(\vec{p}[3]{=}1 \mid M)$$
$$= 1 \times 0.6 \times 0.7 = 0.42$$

$$\Rightarrow Pr(\vec{p}_1 \mid U) = Pr(\vec{p}[1]{=}1 \mid U) \times Pr(\vec{p}[2]{=}0 \mid U) \times Pr(\vec{p}[3]{=}1 \mid U)$$
$$= 0.1 \times 0.5 \times 0.02 = 0.001$$

$$\Rightarrow \frac{Pr(\vec{p}_1|M)}{Pr(\vec{p}_1|U)} = \frac{0.42}{0.001} = 420$$

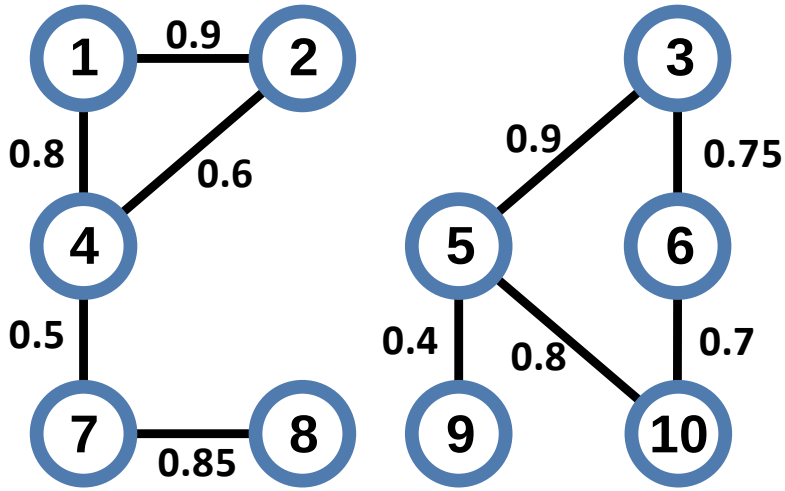
Satz 4: Auf jeden Match kommen genau 499 Unmatches

$$\Rightarrow Pr(M) = 0.002 \text{ und } Pr(U) = 0.998$$

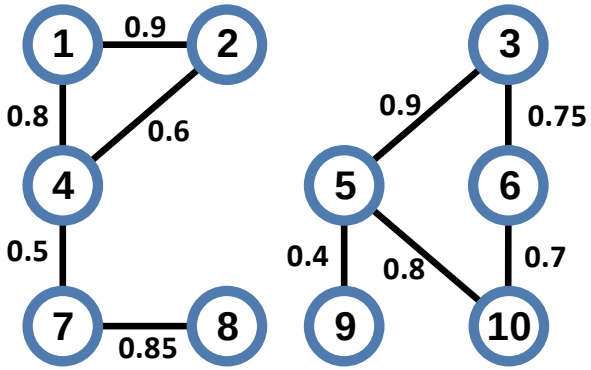
$$\Rightarrow \frac{Pr(U)}{Pr(M)} = \frac{0.998}{0.002} = 499$$

$\Rightarrow$ Ein Tupelpaar welches durch das Pattern $\vec{p}_1 = \langle 1, 0, 1 \rangle$ repräsentiert wird, wird als Unmatch klassifiziert

# Aufgabe 6: Partitioning based on Centers



# Aufgabe 6: Partitioning based on Centers




| | | | |
|---|---|---|---|
| $\{1,2\}$ | Fall 4 | 1 neuer Center | $\{\langle 1,2\rangle\}$ |
| $\{3,5\}$ | Fall 4 | 3 neuer Center | $\{\langle 1,2\rangle,\langle 3,5\rangle\}$ |
| $\{7,8\}$ | Fall 4 | 7 neuer Center | $\{\langle 1,2\rangle,\langle 3,5\rangle,\langle 7,8\rangle\}$ |
| $\{1,4\}$ | Fall 3 | - | $\{\langle 1,2,4\rangle,\langle 3,5\rangle,\langle 7,8\rangle\}$ |
| $\{5,10\}$ | Fall 2 | 10 neuer Center | $\{\langle 1,2,4\rangle,\langle 3,5\rangle,\langle 7,8\rangle,\langle 10\rangle\}$ |
| $\{3,6\}$ | Fall 3 | - | $\{\langle 1,2,4\rangle,\langle 3,5,6\rangle,\langle 7,8\rangle,\langle 10\rangle\}$ |
| $\{6,10\}$ | Fall 1 | - | $\{\langle 1,2,4\rangle,\langle 3,5,6\rangle,\langle 7,8\rangle,\langle 10\rangle\}$ |
| $\{2,4\}$ | Fall 1 | - | $\{\langle 1,2,4\rangle,\langle 3,5,6\rangle,\langle 7,8\rangle,\langle 10\rangle\}$ |
| $\{4,7\}$ | Fall 1 | - | $\{\langle 1,2,4\rangle,\langle 3,5,6\rangle,\langle 7,8\rangle,\langle 10\rangle\}$ |
| $\{5,9\}$ | Fall 2 | 9 neuer Center | $\{\langle 1,2,4\rangle,\langle 3,5,6\rangle,\langle 7,8\rangle,\langle 9\rangle,\langle 10\rangle\}$ |