Stochastisches Matchen

Klaus Pommerening
IMBEI
Oberseminar Medizin-Informatik
16. 7. 2002

Entscheidungssituation

$Y = M \cup U$ disjunkte Zerlegung

γ Beobachtungsfunktion

Γ Merkmalsraum

τ Entscheidungsfunktion

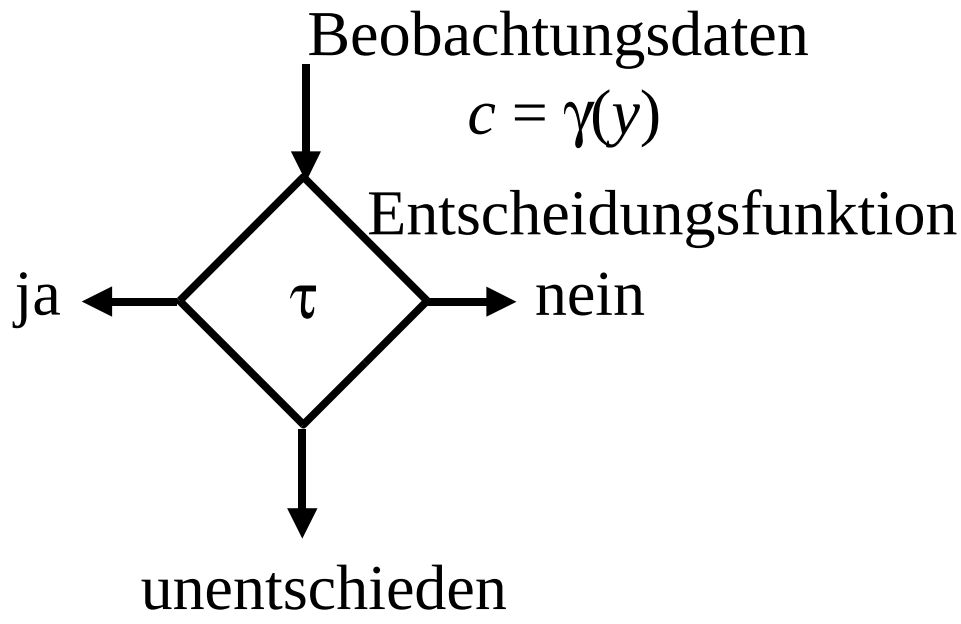
E = {ja, unentschieden, nein}

Alle Mengen endlich.

Aufgabe: Finde τ so, dass möglichst

- $\tau(\gamma(y))$ = ja, wenn $y \in M$,
- $\tau(\gamma(y))$ = nein, wenn $y \in U$.

[Entscheidung zwischen M und U aufgrund der Beobachtung $\gamma(y)$.]



Beispiel (GPOH-PID-Dienst)

- Pflichtdaten (Name, Vorname, Geburtsdatum),
- optionale Daten (z. B. Wohnort).
- Entscheidungssituation z. B. $S_0X_*O_0$:
  - unsichere Daten (S_0),
  - Prüfung auf Übereinstimmung oder Ähnlichkeit (X_*),
  - ohne optionale Daten (O_0).
- Mögliche Ergebnisse – passen zwei Datensätze?
  - ja = Pflichtdaten stimmen exakt überein,
  - unentschieden = Pflichtdaten ähnlich (i. w. Phonetik),
  - nein = Pflichtdaten nicht ähnlich.

Entscheidungsfehler

- **Fehler 1. Art** (falsches Ja): $y \in U$ mit $\tau(\gamma(y)) =$ ja.
  - Fehlerrate 1. Art: $\alpha(\tau) = P(\text{ja}|U) = \Sigma_{c\in\Gamma}\, u(c)\, \delta_{\tau(c),\text{ja}}$
  - Gewichtsfaktor $u(c) = P(c|U)$.

- **Fehler 2. Art** (falsches Nein): $y \in M$ mit $\tau(\gamma(y)) =$ nein.
  - Fehlerrate 2. Art: $\alpha(\tau) = P(\text{nein}|M) = \Sigma_{c\in\Gamma}\, m(c)\, \delta_{\tau(c),\text{nein}}$
  - Gewichtsfaktor $m(c) = P(c|M)$.

- **Unentschiedenheitsrate:**
  - $\eta(\tau) = P(\text{un.}) = \Sigma_{c\in\Gamma}\, n(c)\, \delta_{\tau(c),\text{un.}}$
  - Gewichtsfaktor $n(c) = P(c)$.

Optimierungsziel

- Finde eine Entscheidungsfunktion τ, die
  - bei gegebenen Schranken für die Fehlerraten 1. und 2. Art
  - die Unentschiedenheitsrate minimiert.

[Andere Optimierungsziele denkbar. Bsp. statistische Testsituation: Minimiere Fehlerrate 2. Art bei Schranke für Fehlerrate 1. Art und Unentschiedenheitsrate 0.]

Randomisierte Entscheidungen

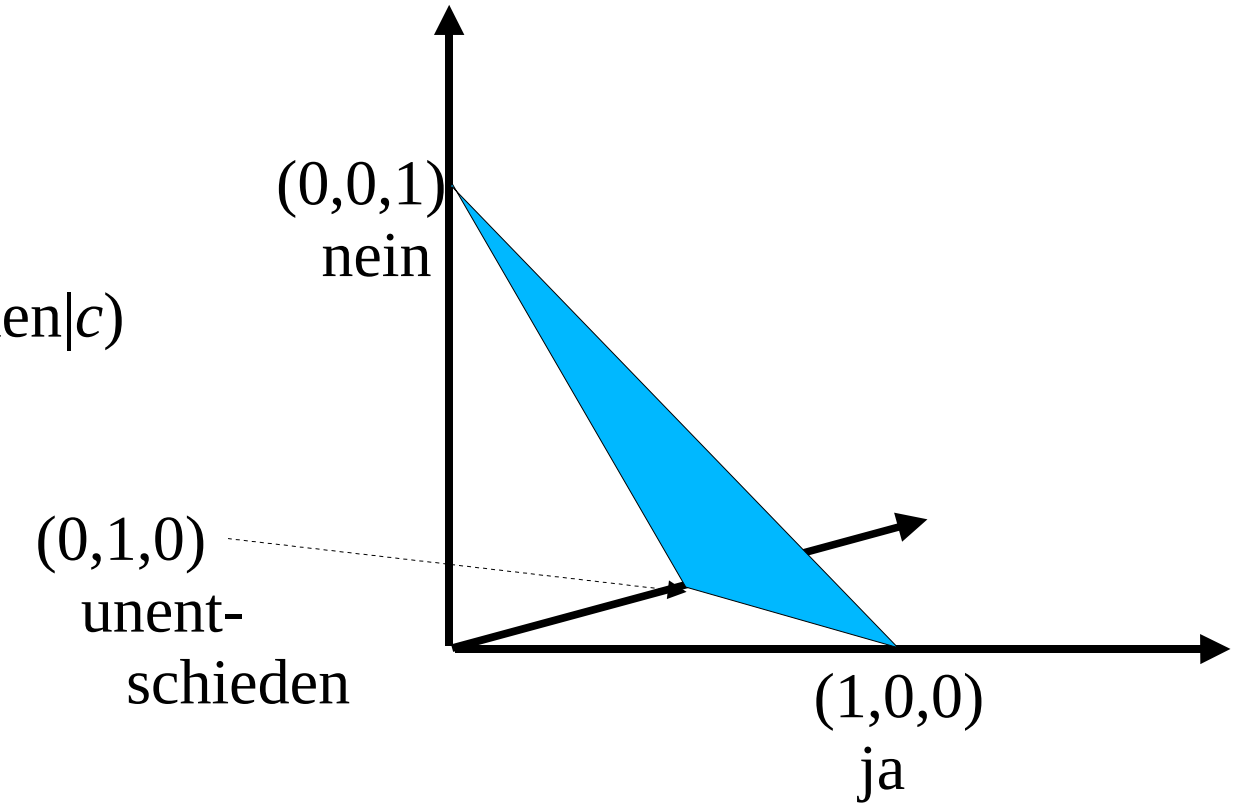
Stochastische Entscheidungsfunktion:

$\tau: \Gamma \longrightarrow [0,1]^3$ mit $\tau_1 + \tau_2 + \tau_3 = 1$ konstant.

$\tau_1(c) = P(\text{ja}|c)$

$\tau_2(c) = P(\text{unentschieden}|c)$

$\tau_3(c) = P(\text{nein}|c)$



Fehlerraten

Unentschiedenheitsrate:

$$\eta(\tau) = P(\text{un.}) = \Sigma_{c\in\Gamma}\, P(c)\, P(\text{un.}|c) = \Sigma_{c\in\Gamma}\, n(c)\, \tau_2(c)$$

Fehlerrate 1. Art:

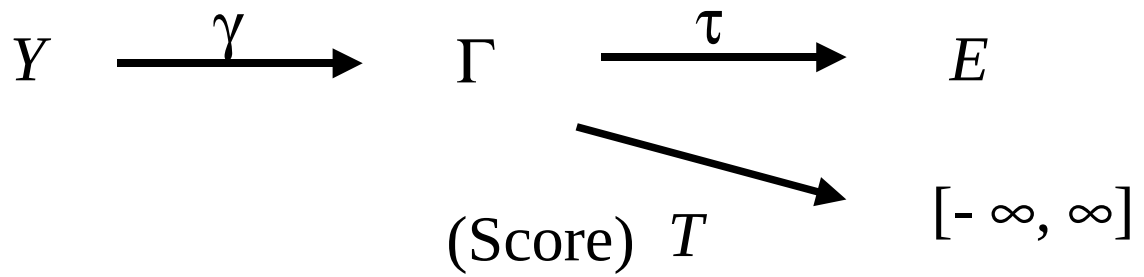
$$\alpha(\tau) = P(\text{ja}|U) = \Sigma_{c\in\Gamma}\, P(c|U)\, P(\text{ja}|c) = \Sigma_{c\in\Gamma}\, u(c)\, \tau_1(c)$$

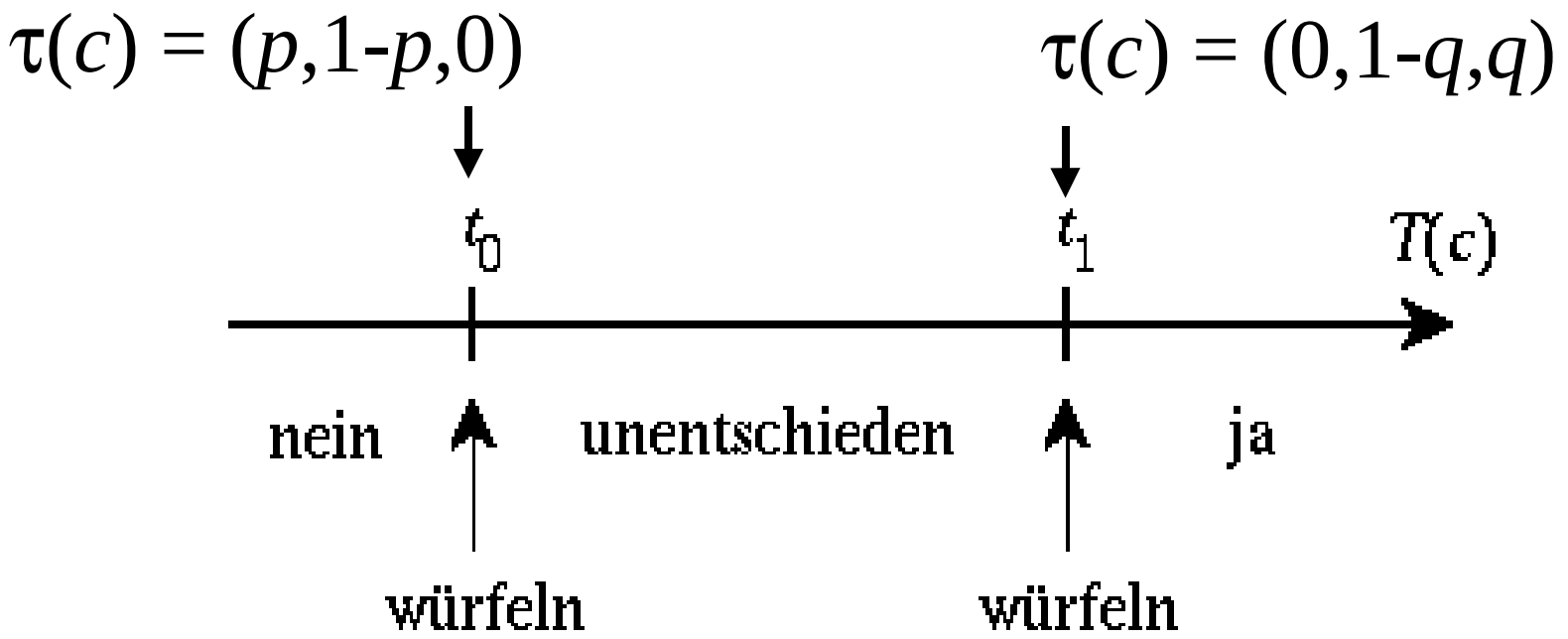
Fehlerrate 2. Art:

$$\beta(\tau) = \text{P}(\text{nein}|M) = \Sigma_{c\in\Gamma}\, P(c|M)\, P(\text{nein}|c)$$

$$= \Sigma_{c\in\Gamma}\, m(c)\, \tau_3(c)$$

Schwellenwert-Verfahren





Die Gewichtsfunktion

Das (i. w.) optimale stochastische Entscheidungsverfahren ist das Schwellenwert-Verfahren zur **Gewichtsfunktion**

$$w := \log m/u : \Gamma \longrightarrow [-\infty, \infty]$$

$$\text{D. h.} \quad w(c) = \log \frac{P(c|M)}{P(c|U)} \qquad (\text{»Odds«}).$$

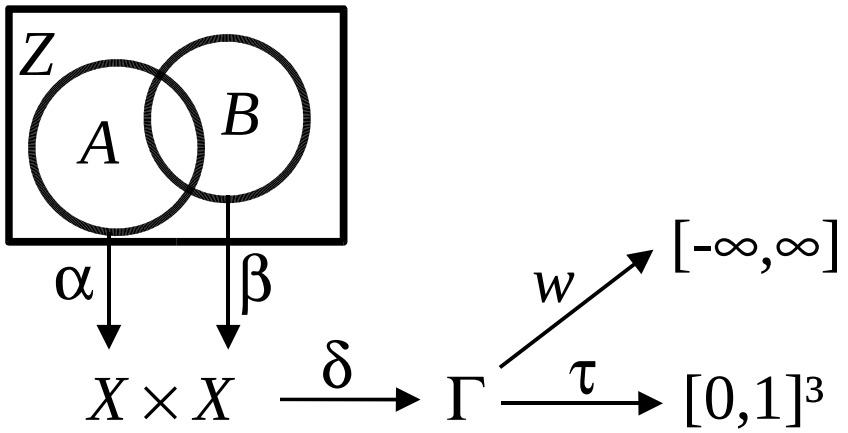
Falls $\Gamma = \Gamma_1 \times \cdots \times \Gamma_K$ (direktes Produkt, Gruppen unabhängiger Merkmale), wird die Gewichtsfunktion entsprechend additiv zerlegt.

Das Match-Problem

- Grundmenge Z (gedachte Population).
- Teilmengen $A, B \subseteq Z$ (mögliche Überschneidung).
- $M = \{(a,b) \in A \times B \mid a = b\}$ (»Matches«),
- $U = \{(a,b) \in A \times B \mid a \neq b\}$ (»unpassende Paare«).
- $Y = A \times B = M \cup U$ (disjunkt).
- Datenerzeugende Prozesse (fehlerbehaftet)

  $\alpha: A \to X, \qquad \beta: B \to X.$
- Entscheidungsproblem = Match-Problem.

Die Match-Situation



Vergleichsfunktion

$\delta: X \times X \to \Gamma.$

(z. B. $\Gamma = X \times X$, δ = identische Abb. oder Γ = {gleich, ungleich}).

- Beobachtungsfunktion $\gamma = \delta \circ (\alpha,\beta): A \times B \to \Gamma.$

  - **Homonymfehler** = Fehler 1. Art,
  - **Synonymfehler** = Fehler 2. Art.

Bestimmung der Gewichte

(für jeweils unabhängige Merkmalsblöcke)

- Im allgemeinen nicht verfügbar, bestenfalls schätzbar.
- Beispiel: Komponente X des Merkmalsraums mit bekannten Wahrscheinlichkeiten $p(x)$, $x \in X$, für das Vorkommen in Z, A, B, $A \cap B$.
- Zwei typische Fälle:
  - Prüfung auf Übereinstimmung: $\Gamma = \{\text{gleich, ungleich}\}$.
  - Prüfung anhand der Werte: $\Gamma = X \times X$.

Prüfung auf Übereinstimmung

$$u(\text{gleich}) = P(\alpha(a) = \beta(b) \mid a \in A, b \in B, a \neq b)$$

$$= \Sigma_{x \in X} P(a \neq b, \alpha(a) = x, \beta(b) = x) \approx \Sigma_{x \in X} p(x)^2$$

$$u(\text{ungleich}) \approx 1 - \Sigma_{x \in X} p(x)^2$$

$$m(\text{gleich}) = P(\alpha(a) = \beta(b) \mid a \in A \cap B)$$

$$\approx (1 - \varepsilon_A)(1 - \varepsilon_B) \quad \text{[Eingabefehler, klein]}$$

$$m(\text{ungleich}) \approx \varepsilon_A + \varepsilon_B - \varepsilon_A \varepsilon_B$$

Zähler: Eingabefehler – **Nenner:** Wahrscheinlichkeiten

Achtung: Änderungen (z. B. im Namen) unter »Eingabefehler«

Beispiele

X gleichverteilt mit $n = \#X$, $\varepsilon = \varepsilon_A = \varepsilon_B$.

- w(gleich) ≈ log $n(1-2\varepsilon)$ (linear in log n)
- w(ungleich) ≈ log $[2n\varepsilon/(n-1)]$ (von n unabhängig)
- **Beispiel 1:** $X = \{\spadesuit, \heartsuit\}$, $n = 2$, $\varepsilon = 5\%$
  - w(gleich) ≈ log 1.8 ≈ 0.26
  - w(ungleich) ≈ log 0.2 ≈ -0.70
- **Beispiel 2:** X = {Jan, ..., Dez}, $n = 12$, $\varepsilon = 5\%$
  - w(gleich) ≈ log 10.8 ≈ 1.03
  - w(ungleich) ≈ log 0.11 ≈ -0.96

Prüfung anhand der Werte

Ohne Beweis:

$$w(x,x) \approx \log \frac{1 - \varepsilon_A + \varepsilon_B}{p(x)}$$

$$w(x,x') \approx \log \frac{p(x)\varepsilon_B + p(x')\varepsilon_A}{(n-1)p(x)p(x')} \quad \text{für } x \neq x'$$

Bei Gleichverteilung: $w(x,x) = w(\text{gleich})$

$w(x,x') = w(\text{ungleich})$

D. h. keine Änderung durch Berücksichtung der Werte.

Beispiel: Merkmal mit zwei Ausprägungen

$X = \{x,x'\}$, $p(x) = q$, $p(x') = 1\text{-}q$

z. B. $\varepsilon = 10\%$, $q = 1/4$

- w(gleich) ≈ log 1.28 ≈ 0.11
- w(ungleich) ≈ log 0.53 ≈ -0.27
- $w(x,x)$ ≈ log 3.2 ≈ 0.51 [erhöht!]
- $w(x',x')$ ≈ log 1.07 ≈ 0.03 [erniedrigt!]
- w(x,x') ≈ log 0.53 ≈ -0.27

Typische Merkmale I

- Geschlecht: 2 Ausprägungen (ohne fehlende Werte), gleichverteilt:
  - Unterscheidung gleich/ungleich reicht,
  - zu schätzender Parameter nur ε.
- Geburtsmonat: analog mit $n = 12$.
- Geburtsjahr: *nicht* gleichverteilt:
  - Verteilung und Fehlerraten zu schätzen.
- Vorname: »leicht« abhängig von Name und Geburtsjahr.

Typische Merkmale II

- Name: komplexes Modell für Beobachtungsfunktion
  - Merkmalsraum X = Zeichenketten der Länge $\leq N$.
  - 1. Schritt: Normalisierung Name $\rightarrow$ (N_1, N_2, N_3).
  - 2. Schritt: Phonetik Ph_K (Köln), Ph_H (Hannover).
  - $\Gamma = X^{10}$.
  - γ(Name, Name') = (N_1,N_2,N_3,Ph_K(Name), Ph_H(Name), N_1',...).
  - Unterschiedliche Gewichte für Gleichheit, Ähnlichkeit in einem oder beiden phonetischen Codes.

Algorithmus

- Deterministische Vorentscheidung („Blockbildung“)
  - Stochastische Entscheidungen einbetten
- Gruppen unabhängiger Variablen definieren
  - Welches Prüfverfahren für welche Merkmale?
  - Komplizierte Merkmale: Namen, ... - wie behandeln?
- Eingabefehler woher schätzen?
  - Startwerte, dynamisches Update?
- Wahrscheinlichkeiten aus Datenbank schätzen!
  - Startwerte? Dynamisches Update!

Fragen

- Passt die Theorie auf das Szenario:
  - Jeweils 1 neuer Fall wird mit der Datenbank abgeglichen und ggf. neu hinzugefügt?
- Besondere Behandlung fehlender Werte?
- Wie geht man mit echten Homonymen um?
- Wie wirkt sich eine Fehlschätzung der Eingabefehlerraten aus?

Weiterführende Probleme I

- Optimierung des algorithmischen Ablaufs des Match-Verfahrens
  - Optimierung abhängig von der Vergleichsfunktion?
  - Stabilität des Verfahrens bei Schwankungen der geschätzten Parameter? Geht Optimalität verloren?
  - Folgen der Verletzung von Unabhängigkeitsannahmen?
- Adjustierung der Gewichte – genetischer Algorithmus?
- Wahl und Adjustierung der Schwellenwerte?
- Optimierung des Match-Verfahrens nach anderen Kriterien? (Z. B. Minimierung der HFR, Balance zwischen HFR/SFR?

Weiterführende Probleme II

- Andere Ansätze für das Match-Verfahren? Varianten? (Literatur!)
- Testweiser Parallelbetrieb von stochastischem und deterministischen Matchen mit Vergleich der Ergebnisse.