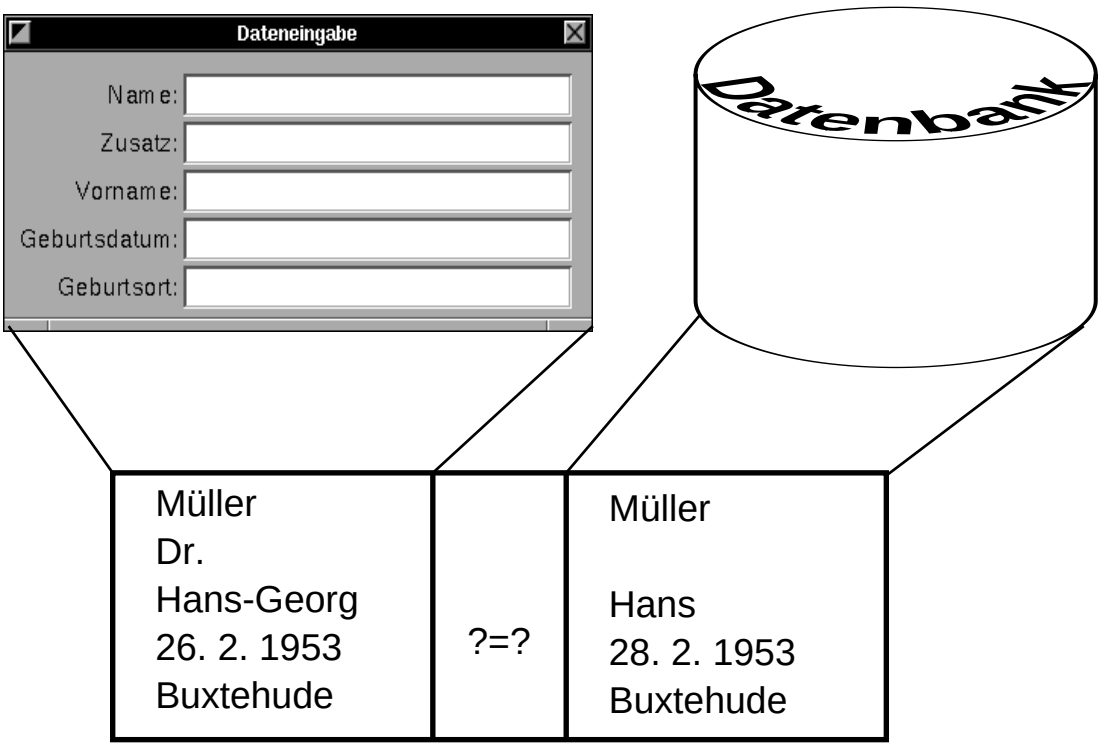
Stochastisches Matchen

Klaus Pommerening
Institut für Medizinische Biometrie, Epidemiologie und Informatik
der Johannes-Gutenberg-Universität
D-55101 Mainz

Unvollständiger Entwurf
28. Mai 2002 – letzte Änderung: 1. Juni 2003



Einleitung

Beim Aufbau von krankheitsbezogenen Registern (oder Patientenlisten) im Rahmen medizinischer Forschungsverbünde tritt das Problem auf, das in der Epidemiologie schon lange als Match- oder Record-Linkage-Problem bekannt ist: Wie erkennt man, ob ein neu gemeldeter Fall – der eventuell mit Fehlern, wie einem falsch geschriebenen Namen, behaftet ist, – einem bereits registrierten Fall entspricht? Als Randbedingung hinzu kommen ärztliche Schweigepflicht und Datenschutz, die meist erfordern, dass Meldestelle und Registerstelle institutionell getrennt sind, erstere keine medizinischen Daten erhält oder speichert und letztere keine Identitätsdaten, sondern statt dessen Pseudonyme.

Die Zuordnung, das „Matchen“, soll dabei weitgehend automatisch bei der Meldung eines Falles ablaufen. Fehler in der Zuordnung sind dabei nicht mit völliger Sicherheit auszuschließen, wie allerdings auch bei einem von Menschen durchgeführten Abgleichvorgang.

Unter diesen Anforderungen soll ein Verfahren konzipiert werden, das möglichst wenig Fehler macht oder wenigstens bestimmte Fehlerquoten nicht überschreitet. Hierfür wurde ein deterministischer Algorithmus entwickelt und im PID-Dienst der TMF realisiert, der zufriedenstellend funktioniert. Etwas bessere Ergebnisse sollte man aber noch erhalten können, wenn man die Verfahren zum stochastischen Matchen verwendet, die in verschiedenen Versionen auf Newcombe, Fellegi/Sunter, Jaro u. a. zurückgehen und in dem Software-Paket AUTOMATCH angeboten werden.

Diese Verfahren zum stochastischen Matchen sollen hier in einer für den Aufbau von Registern geeigneten Form beschrieben werden.

Für Mathematiker sicherlich unbefriedigend sind zwei Aspekte dieses Artikels:

- Der naive Umgang mit Wahrscheinlichkeiten; diese werden ohne Bedenken mit relativen Häufigkeiten gleichgesetzt, anstatt eine stochastisch saubere Situation mit Wahrscheinlichkeitsraum, Zufallsvariablen und Stichproben zu betrachten.

- Der großzügige Umgang mit Näherungsformeln ohne den Versuch, Fehlerschranken herzuleiten.

1 Optimale dreiwertige Entscheidungen

Informeller Überblick

An Objekten y aus einer Menge Y werden Messungen oder Beobachtungen von K unabhängigen Merkmalen vorgenommen – es wird dabei ein Satz c von Werten $c_1, \ldots, c_K$ gewonnen. Aufgrund dieser Werte soll entschieden werden, ob y zu einer bestimmten Teilmenge $M \subseteq Y$ (Entscheidung „ja“) oder zu ihrem Komplement $U = Y - M$ gehört (Entscheidung „nein“). In Zweifelsfällen ist auch die Entscheidung „unentschieden“ möglich.

Es werden Entscheidungsverfahren in Abhängigkeit von dem Messwert-Satz c gesucht, die vorgegebe Obergrenzen für die beiden Arten von Fehlern – fälschliche Entscheidung „ja“, obwohl $y \in M$, und fälschliche Entscheidung „nein“, obwohl $y \in U$, – einhalten; unter all diesen Verfahren soll eines gefunden werden, das die Anzahl der „unentschiedenen“ Ausgänge minimiert.

Die Lösung sieht so aus: Für jeden der Werte c_i sei bekannt, mit welcher Wahrscheinlichkeit er in M bzw. U auftritt, informell geschrieben als $P(c_i|M)$ bzw. $P(c_i|U)$. Daraus werden die Quotienten („Likelihood Ratios“)

$$P(c_i|M)/P(c_i|U)$$

gebildet – ein größerer Wert eines solchen Quotienten spricht eher für M, ein kleinerer eher für U. Diese Quotienten werden miteinander multipliziert; der Logarithmus des Produkts wird als „Gewichtsfunktion“ w bezeichnet:

$$w(c) = \log \frac{P(c_1|M) \cdots P(c_K|M)}{P(c_1|U) \cdots P(c_K|U)} = \log \frac{P(c_1|M)}{P(c_1|U)} + \cdots + \log \frac{P(c_K|M)}{P(c_K|U)}.$$

Ferner werden Schwellenwerte t_0 und t_1 festgelegt, und entschieden wird:

$$\begin{cases} \text{ja}, & \text{falls } w(c) > t_1, \\ \text{unentschieden}, & \text{falls } t_0 < w(c) < t_1, \\ \text{nein}, & \text{falls } T(c) < t_0. \end{cases}$$

Falls gerade genau einer Schwellenwerte getroffen wird, randomisiert man die Entscheidung.

In diesem Kapitel wird dieses Ergebnis mathematisch hergeleitet.

1.1 Abstrakte Entscheidungssituation

Wir stellen uns eine endliche Grundmenge Y vor, die aus zwei nichtleeren disjunkten Teilmengen M und U besteht:

$$Y = M \dot\cup U.$$

Wir stellen uns weiter vor, dass wir gewisse Beobachtungen oder Messungen an einem Element $y \in Y$ durchführen können und danach entscheiden

müssen, ob $y \in M$ oder $y \in U$. Zur Beschreibung dieser Situation nehmen wir in unser Modell einen „Merkmalsraum“ Γ und eine „Beobachtungsfunktion“

$$\gamma\colon Y \longrightarrow \Gamma$$

auf, die auch fehlerbehaftet sein kann. Das Entscheidungsverfahren modellieren wir zunächst durch eine Funktion

$$\tau\colon \Gamma \longrightarrow E$$

in eine dreielementige Menge von „Entscheidungen“, deren Elemente als „ja“ (= Entscheidung für M), „unentschieden“ und „nein“ (= Entscheidung für U) bezeichnet werden.

> **Vorstellung:** Es soll „nach Möglichkeit“ $\tau \circ \gamma(y) =$ ja sein, wenn $y \in M$, und = nein, wenn $y \in U$.

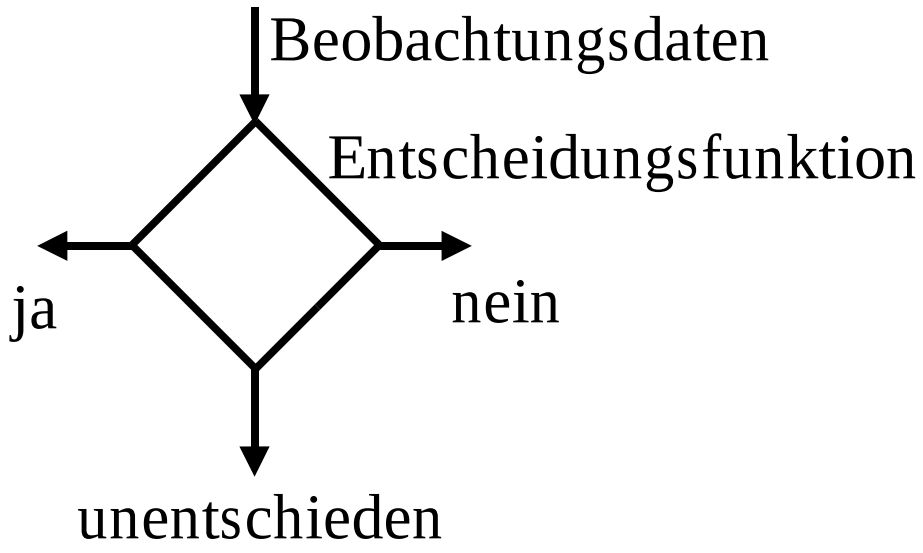


Abbildung 1: Die Entscheidungsraute

Beispiele für solche Situationen sind:

- Klassifikationsprobleme, z. B. in der medizinischen Diagnostik die Unterscheidung zwischen einem positiven und einem negativen Befund.
- Die Testsituation in der Statistik: soll die „Nullhypothese“ verworfen (= positives Testergebnis) oder angenommen (= negatives Testergebnis) werden? (Hier wird das Ergebnis „unentschieden“ in der Regel nicht zugelassen.)

- Das Match-Problem, siehe Abschnitt 2: Gehören zwei auf verschiedene Weise gewonnenen Datensätze zum gleichen Individuum? Ist der neu in die Datenbank eingegebene Fall schon früher registriert worden?

Beispiel. Beim GPOH-PID-Dienst wird unterschieden zwischen „Pflichtdaten“ (Name, Vorname, Geburtsdatum) und „optionalen Daten“ (u. a. Wohnort). Die Entscheidungssituationen beim Matchen sind durch Kürzel bezeichnet. So bedeutet beispielsweise $S_0X_*O_0$:

- S_0: Mindestens einer der beiden zu vergleichenden Datensätze ist nicht als „sicher“ gekennzeichnet, d. h., er kann Fehler enthalten.
- X_*: Die Pflichtdaten werden auf exakte Übereinstimmung (X_1) oder Ähnlichkeit (X_0) überprüft. Dabei werden erkennbare Namensänderungen oder Vertauschung der Reihenfolge zweier Vornamen als exakte Übereinstimmung gewertet; Ähnlichkeit bedeutet im wesentlichen phonetische Übereinstimmung.
- O_0: Mindestens einer der beiden Datensätze enthält keine optionalen Daten, oder die ausgefüllten optionalen Datenfelder sind disjunkt.

Die möglichen Ergebnisse sind:

- ja (die Datensätze gehören zu einer Person), wenn die Pflichtdaten exakt übereinstimmen,
- unentschieden (eventuell einem weiteren Entscheidungsverfahren zuzuführen oder entsprechende Rückmeldung), wenn die Pflichtdaten ähnlich sind,
- nein (die Datensätze gehören nicht zur selben Person), wenn die Pflichtdaten nicht einmal ähnlich sind.

Ein Entscheidungsverfahren ist gut, wenn es möglichst wenige Fehlentscheidungen oder möglichst wenige unentschiedene Situationen liefert. Diese beiden Zielkriterien sind nicht unabhängig voneinander optimierbar: Ein Verfahren, das immer „unentschieden“ antwortet, macht keine Fehler, nützt aber nichts. Ein Verfahren, das einfach zwischen „ja“ und „nein“ würfelt, lässt keinen Fall unentschieden, ist aber genau so nutzlos. Sinnvolle Optimierungen sind Thema der nächsten Abschnitte.

1.2 Entscheidungsfehler

Entschieden werden soll also aufgrund der zusammengesetzten Abbildung

$$Y \xrightarrow{\gamma} \Gamma \xrightarrow{\tau} E.$$

Wir setzen von jetzt an voraus, dass auch der Merkmalsraum Γ, genau wie die Grundmenge Y, eine endliche Menge ist.

Ein **Fehler erster Art** ist ein Element $y \in U$ mit $\tau \circ \gamma(y) = \text{ja}$.

> **Zielvorstellung:** Die Zahl der Fehler erster Art des Entscheidungsverfahrens τ soll möglichst klein sein oder wenigstens einen vorgegebenen Wert nicht überschreiten.
>
> In der Diagnostik ist dies das fälschliche Beobachten eines positiven Testergebnisses, in der statistischen Testsituation die fälschliche „Verwerfung der Nullhypothese“, beim Match-Problem das fälschliche Zusammenführen, der „Homonymfehler“.

Die Menge der Fehler erster Art ist

$$\begin{aligned}\{y \in U \mid \tau \circ \gamma(y) = \text{ja}\} &= \dot{\bigcup_{c\in\Gamma,\tau(c)=\text{ja}}} \{y \in U \mid \gamma(y) = c\} \\ &= \dot{\bigcup_{c\in\Gamma,\tau(c)=\text{ja}}} U \cap \gamma^{-1}(c),\end{aligned}$$

ihre Größe also

$$\begin{aligned}\#\{y \in U \mid \tau \circ \gamma(y) = \text{ja}\} &= \sum_{c\in\Gamma,\tau(c)=\text{ja}} \#(U \cap \gamma^{-1}(c)) \\ &= \sum_{c\in\Gamma} \#(U \cap \gamma^{-1}(c)) \cdot \delta_{\tau(c),\text{ja}}\end{aligned}$$

mit dem KRONECKER-Symbol δ. Ihre relative Größe, die „Fehlerrate erster Art“ oder „Wahrscheinlichkeit eines Fehlers erster Art“, ist daher

$$\alpha(\tau) = \frac{\#\{y \in U \mid \tau \circ \gamma(y) = \text{ja}\}}{\#U} = \sum_{c\in\Gamma} u(c) \cdot \delta_{\tau(c),\text{ja}}$$

mit dem Gewichtsfaktor

$$u(c) := \frac{\#(U \cap \gamma^{-1}(c))}{\#U},$$

der die Wahrscheinlichkeit angibt, mit der ein Element von U durch γ auf den „Beobachtungsdatensatz“ $c \in \Gamma$ abgebildet wird; als bedingte Wahrscheinlichkeit würde man $u(c)$ suggestiv, aber formal unvollkommen, als $P(c|U)$ schreiben.

Ein **Fehler zweiter Art** ist ein Element $y \in M$ mit $\tau \circ \gamma(y) = \text{nein}$.

> **Zielvorstellung:** Auch die Zahl der Fehler zweiter Art des Entscheidungsverfahrens τ soll möglichst klein sein oder wenigstens einen vorgegebenen Wert nicht überschreiten.

In der Diagnostik ist dies das fälschliche Beobachten eines negativen Testergebnisses, in der statistischen Testsituation die fälschliche „Annahme der Nullhypothese“, beim Match-Problem das fälschliche Nicht-Zusammenführen, der „Synonymfehler“.

Die Menge der Fehler zweiter Art ist

$$\begin{aligned}\{y \in M \mid \tau \circ \gamma(y) = \text{nein}\} &= \dot{\bigcup_{c\in\Gamma,\tau(c)=\text{nein}}} \{y \in M \mid \gamma(a,b) = c\} \\ &= \dot{\bigcup_{c\in\Gamma,\tau(c)=\text{nein}}} M \cap \gamma^{-1}(c),\end{aligned}$$

ihre Größe also

$$\#\{y \in M \mid \tau \circ \gamma(y) = \text{nein}\} = \sum_{c\in\Gamma} \#(M \cap \gamma^{-1}(c)) \cdot \delta_{\tau(c),\text{nein}},$$

ihre relative Größe, die „Fehlerrate zweiter Art“ oder „Wahrscheinlichkeit eines Fehlers zweiter Art“, ist daher

$$\beta(\tau) = \frac{\#\{y \in M \mid \tau \circ \gamma(y) = \text{nein}\}}{\#M} = \sum_{c\in\Gamma} m(c) \cdot \delta_{\tau(c),\text{nein}}$$

mit dem Gewichtsfaktor

$$m(c) := \frac{\#(M \cap \gamma^{-1}(c))}{\#M},$$

der die Wahrscheinlichkeit angibt, mit der ein Element von M durch γ auf c abgebildet wird und den man als bedingte Wahrscheinlichkeit in der Form $P(c|M)$ schreiben würde.

Ziel. Es soll eine Entscheidungsfunktion $\tau : \Gamma \longrightarrow E$, gefunden werden, die bei vorgegebenen Schranken für Fehler erster und zweiter Art die Zahl der unentschiedenen Fälle minimiert; diese Zahl ist

$$\begin{aligned}\#\{y \in Y \mid \tau \circ \gamma(y) = \text{unentschieden}\} &= \dot{\bigcup_{c\in\Gamma,\tau(c)=\text{un.}}} \{y \in Y \mid \gamma(y) = c\} \\ &= \dot{\bigcup_{c\in\Gamma,\tau(c)=\text{un.}}} \gamma^{-1}(c),\end{aligned}$$

ihre relative Größe also

$$\begin{aligned}\eta(\tau) &= \frac{\#\{y \in Y \mid \tau \circ \gamma(y) = \text{un.}\}}{\#Y} \\ &= \sum_{c\in\Gamma,\tau(c)=\text{un.}} \frac{\#\gamma^{-1}(c)}{\#Y} \\ &= \sum_{c\in\Gamma} n(c) \cdot \delta_{\tau(c),\text{un.}}\end{aligned}$$

mit dem Gewichtsfaktor

$$n(c) = \frac{\#\gamma^{-1}(c)}{\#Y},$$

der die Wahrscheinlichkeit angibt, mit der ein beliebiges Element $\in Y$ auf den Beobachtungsdatensatz c abgebildet wird (und die ein Stochastiker mit $P(c)$ bezeichnen würde).

Damit sind die drei Gewichtsfunktionen

$$m, u, n : \Gamma \longrightarrow [0, 1]$$

definiert. Klar, dass

$$\sum_{c\in\Gamma} m(c) = \sum_{c\in\Gamma} u(c) = \sum_{c\in\Gamma} n(c) = 1.$$

Da Γ und E endliche Mengen sind, ist die gesuchte optimale Entscheidungsfunktion durch vollständige Suche auf jeden Fall zu finden, vorausgesetzt, es gibt überhaupt ein Entscheidungsverfahren, das die verlangten Fehlerschranken einhält. Eine solche vollständige Suche ist allerdings nicht effizient.

Ein anderes Optimierungsziel wird meist in der statistischen Testsituation angestrebt: bei der Unentschiedenheitsrate $\eta(\tau) = 0$ und festgelegter Obergrenze $\alpha(\tau) \leq \alpha_0$ für den Fehler erster Art den Fehler zweiter Art $\beta(\tau)$ zu minimieren, d. h., die „Power“ $1 - \beta(\tau)$ zu maximieren.

1.3 Randomisierte Entscheidungen

Für die Optimierung ist es vorteilhaft, die Entscheidung zwischen „ja“, „unentschieden“ und „nein“ zu randomisieren; „vorteilhaft“ bedeutet, dass es ein besseres Optimum und ein effizientes Verfahren gibt, dieses zu finden. Das führt zu der folgenden Begriffsbildung.

Eine **(stochastische) Entscheidungsfunktion** ist eine Funktion

$$\tau\colon \Gamma \longrightarrow [0, 1]^3 \quad \text{mit } \tau_1 + \tau_2 + \tau_3 = 1 \text{ konstant},$$

die also jedem Vergleichsdatensatz c ein Tripel $\tau_1(c), \tau_2(c), \tau_3(c)$ von Zahlen zuordnet.

> $\tau_1(c)$ bezeichnet die Wahrscheinlichkeit, mit der bei Beobachtung von c für „ja“ entschieden wird, $\tau_2(c)$ die Wahrscheinlichkeit für „unentschieden“, $\tau_3(c)$ die Wahrscheinlichkeit für „nein“. Beim Entscheidungsprozess wird also für jeden Beobachtungsdatensatz c eine dreiseitige Münze geworfen, die mit den Wahrscheinlicheiten $\tau_i(c)$ das jeweilige Ergebnis bringt.

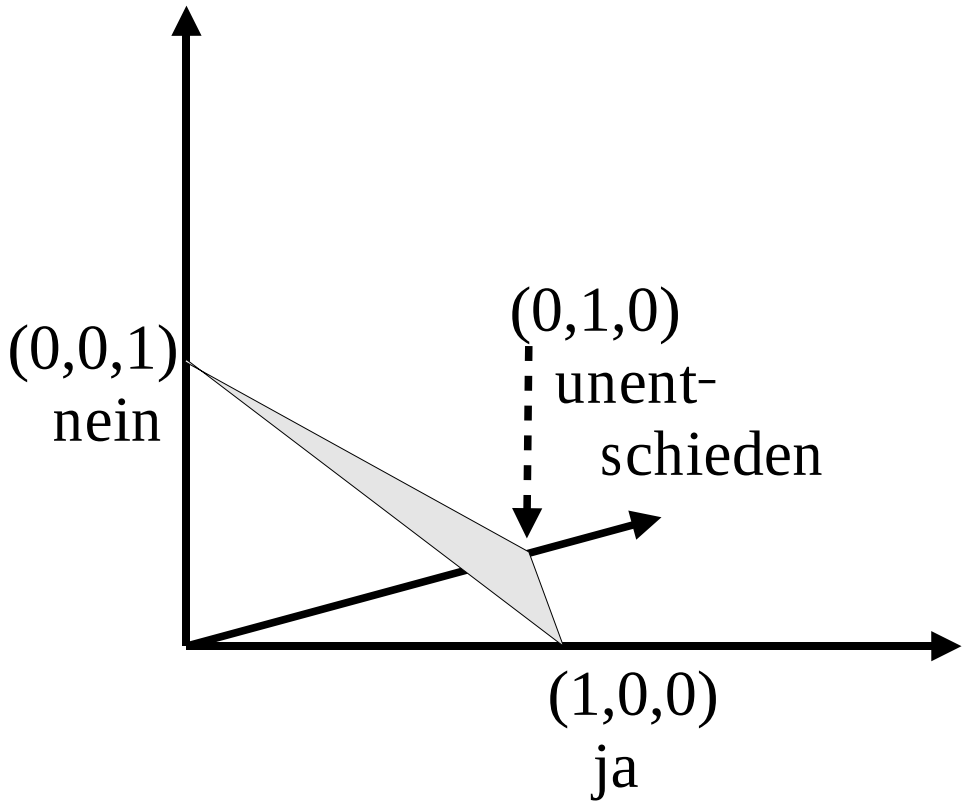


Abbildung 2: Das Dreieck der stochastischen Entscheidungen

Bei einer deterministischen Entscheidung nach Abschnitt 1.2 kamen in den Ausdrücken für die Fehler erster und zweiter Art die Größen $\delta_{\tau(c),x}$ vor, die nur die Werte 0 und 1 annehmen konnten. Ersetzen wir die dortigen Werte „ja“, „unentschieden“ und „nein“ durch $(1,0,0)$, $(0,1,0)$ und $(0,0,1) \in [0,1]^3$, so wird

$$\begin{aligned} \delta_{\tau(c),\text{ja}} &= \delta_{\tau_1(c),1} = \text{Wahrscheinlichkeit, dass } \tau_1(c) = 1 \text{ ist,} \\ \delta_{\tau(c),\text{nein}} &= \delta_{\tau_3(c),1} = \text{Wahrscheinlichkeit, dass } \tau_3(c) = 1 \text{ ist.} \end{aligned}$$

Analog werden daher die **Fehlerrate erster und zweiter Art** des stochastischen Entscheidungsverfahrens τ definiert durch

$$\begin{aligned} \alpha(\tau) &:= \sum_{c\in\Gamma} u(c)\cdot\tau_1(c), \\ \beta(\tau) &:= \sum_{c\in\Gamma} m(c)\cdot\tau_3(c). \end{aligned}$$

Der Summand $u(c)\cdot\tau_1(c)$ ist die Wahrscheinlichkeit, dass ein $y \in U$ auf c abgebildet wird *und* dass dann für „ja“ entschieden wird. Analog für $m(c)\cdot\tau_3(c)$. Also ist die Fehlerrate erster Art die Wahrscheinlichkeit, mit der ein Fehler erster Art begangen wird, und die Fehlerrate zweiter Art die Wahrscheinlichkeit, mit der ein Fehler zweiter Art begangen wird. In der suggestiven Schreibweise der Stochastik sehen die Formeln wahrscheinlich so aus:

$$\begin{aligned} \alpha(\tau) &= \sum_{c\in\Gamma} P(c|U)\cdot P(\text{ja}|c) = P(\text{ja}|U), \\ \beta(\tau) &= \sum_{c\in\Gamma} P(c|M)\cdot P(\text{nein}|c) = P(\text{nein}|M). \end{aligned}$$

Das Optimierungsziel wird jetzt auf stochastische Entscheidungsverfahren ausgedehnt: Minimiert werden soll bei vorgegebenen Schranken

$$\alpha(\tau) \le \mu, \quad \beta(\tau) \le \lambda$$

die Größe („Unentschiedenheitsrate“)

$$\eta(\tau) := \sum_{c\in\Gamma} n(c)\cdot\tau_2(c),$$

die die Wahrscheinlichkeit dafür angibt, dass für ein Element $y \in Y$ die Entscheidung „unentschieden“ getroffen wird; ein Stochastiker würde wahrscheinlich so schreiben:

$$\eta(\tau) = \sum_{c\in\Gamma} P(c)\cdot P(\text{un}|c) = P(\text{un}).$$

Der Funktionenraum $\mathcal{T}$, auf dem optimiert werden soll, besteht also aus den Funktionen $\Gamma \longrightarrow [0,1]^3$. Bei vorgebenen oberen Schranken für

$$\alpha, \beta : \mathcal{T} \longrightarrow [0,1]$$

soll ein Minimum für

$$\eta : \mathcal{T} \longrightarrow [0,1]$$

gefunden werden. Es ist nicht ohne weiteres sicher, dass die Bedingungen $\alpha(\tau) \leq \mu$ und $\beta(\tau) \leq \lambda$ kompatibel sind, d. h., gleichzeitig erfüllt werden können; insbesondere können die Fehlerschranken nicht beide 0 sein, wenn der Beobachtungsprozess γ nicht die Mengen M und U vollständig trennt.

1.4 Schwellenwert-Verfahren

Eine typische stochastische Entscheidungsfunktion $\tau : \Gamma \longrightarrow [0,1]^3$ ist nicht gar so willkürlich; sie ist etwa eine Treppenfunktion, die nur an den Übergängen randomisiert ist. Genauer ist damit folgendes gemeint: Gegeben sei eine Funktion

$$T \colon \Gamma \longrightarrow \bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\},$$

unter der wir uns einen aus dem Beobachtungsdatensatz c abgeleiteten Hilfswert („Score“) vorstellen, und $t_0, t_1 \in \mathbb{R}$ seien Schwellenwerte mit $t_0 \geq t_1$. Ferner seien $p, q \in [0,1]$. Dann ist eine stochastische Entscheidungsfunktion gegeben durch:

$$\tau(c) = \begin{cases} (1,0,0) = \text{ja}, & \text{falls } T(c) > t_1, \\ (p, 1-p, 0), & \text{falls } T(c) = t_1, \\ (0,1,0) = \text{un.}, & \text{falls } t_0 < T(c) < t_1, \\ (0, 1-q, q), & \text{falls } T(c) = t_0, \\ (0,0,1) = \text{nein}, & \text{falls } T(c) < t_0. \end{cases}$$

„Gewürfelt“ wird also nur, wenn der Score gerade einen der Schwellenwerte annimmt.

1.5 Minimierung der Unentschiedenheitsrate

Nach [3] wird jetzt ein stochastisches Entscheidungsverfahren definiert. Die Definition verwendet eine von γ abhängige lineare Ordnung auf der Menge

$$\tilde{\Gamma} = \gamma(Y) \subseteq \Gamma$$

der wirklich beobachteten Beobachtungsdatensätze.

Klar ist, dass für $c \in \Gamma$ gilt

$$\begin{aligned} m(c) = 0 &\iff M \cap \gamma^{-1}(c) = \emptyset \iff c \notin \gamma(M), \\ u(c) = 0 &\iff U \cap \gamma^{-1}(c) = \emptyset \iff c \notin \gamma(U), \\ m(c) = u(c) = 0 &\iff c \notin \tilde{\Gamma}. \end{aligned}$$

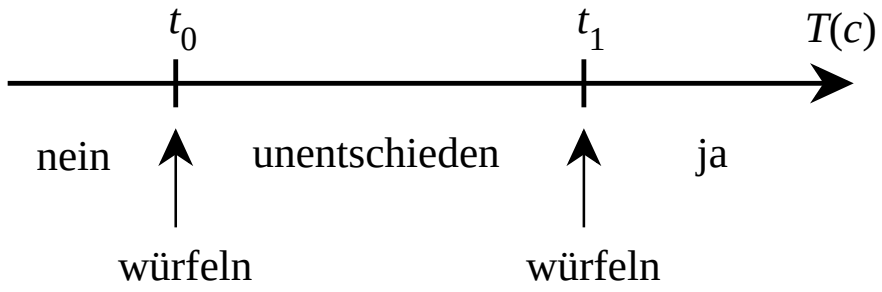


Abbildung 3: Die Schwellenwert-Entscheidung

Für $c \in \tilde{\Gamma}$ ist also mindestens einer der beiden Werte $m(c)$ oder $u(c) > 0$. Daher ist die Quotientenfunktion

$$\frac{m}{u} : \tilde{\Gamma} \longrightarrow \mathbb{R}_+ \cup \{\infty\} = [0, \infty]$$

wohldefiniert.

> Ein größerer Wert von m bedeutet, dass der Beobachtungsdatensatz c auf M häufiger auftritt, ein kleinerer Wert von u, dass c auf U seltener auftritt. Ein größerer Wert des Quotienten heißt also, dass man eher für ja entscheiden sollte, ein kleinerer Wert spricht eher für nein. Der Quotient
>
> $$\frac{m(c)}{u(c)} = \frac{P(c|M)}{P(c|U)}$$
>
> wird als „Likelihood Ratio“, bei NEWCOMBE [7] auch als „Odds“ (= Wettchancen-Verhältnis) bezeichnet; bei Werten größer als 1 wird man für M wetten – also „ja“ entscheiden –, bei Werten kleiner als 1 für U – also „nein“ entscheiden.

Damit wird die lineare Ordnung auf $\tilde{\Gamma}$ so definiert, dass

$$\begin{aligned} c \prec c' &\iff \frac{m}{u}(c) > \frac{m}{u}(c'), \\ c \prec c' &\implies \frac{m}{u}(c) \geq \frac{m}{u}(c'). \end{aligned}$$

D. h., die Elemente von $\tilde{\Gamma}$ werden nach abnehmender Größe des Quotienten $m(c)/u(c)$ angeordnet, wobei bei gleichem Wert des Quotienten irgendeine willkürliche Reihenfolge gewählt wird. Bezüglich dieser Ordnung werden die Elemente von $\tilde{\Gamma}$ durchnummeriert:

$$\tilde{\Gamma} = \{c_1, \ldots, c_N\} \quad \text{mit } c_1 \prec c_2 \prec \ldots \prec c_N.$$

Je größer der Index, desto eher gehört der Beobachtungsdatensatz zu U, je kleiner, desto eher zu M.

Seien nun feste Fehlerschranken $\mu, \lambda \in]0,1]$ gewählt. Dazu findet man einen Index n mit

$$u(c_1) + \cdots + u(c_{n-1}) < \mu \leq u(c_1) + \cdots + u(c_n),$$

denn $u(c_1) + \cdots + u(c_N) = 1$. Da

$$\sum_{i=1}^{n-1} u(c_i) = \frac{\sum_{i=1}^{n-1} \#(U \cap \gamma^{-1}(c_i))}{\#U} = \frac{\#(U \cap \gamma^{-1}\{c_1, \ldots, c_{n-1}\})}{\#U},$$

ist die Fehlerrate erster Art, also die Entscheidung „ja“ für Elemente aus U, so gut ausgeschöpft, wie das deterministisch möglich ist. Für den Index $i = n$ trifft man die Entscheidung durch „Würfeln“ mit Wahrscheinlichkeit p für „ja“ und $1 - p$ für „unentschieden“ so, dass

$$p \cdot u(c_n) = \mu - \sum_{i=1}^{n-1} u(c_i);$$

damit wird die Fehlerrate erster Art zu μ aufgefüllt.

Analog findet man einen Index n' mit

$$m(c_{n'+1}) + \cdots + m(c_N) < \lambda \leq m(c_{n'}) + \cdots + m(c_N).$$

Dazu passend wird die Wahrscheinlichkeit q so gewählt, dass

$$q \cdot m(c_{n'}) = \lambda - \sum_{i=n'+1}^{N} m(c_i).$$

Falls $n < n'$, sind die Fehlerraten μ und λ kompatibel, und das folgende ist ein stochastisches Entscheidungsverfahren:

$$\tau_{\lambda,\mu}(c) := \begin{cases} (1,0,0) & \text{für } 1 \leq i < n, \\ (p, 1-p, 0) & \text{für } i = n, \\ (0,1,0) & \text{für } n < i < n', \\ (0, 1-q, q) & \text{für } i = n', \\ (0,0,1) & \text{für } n' < i \leq N. \end{cases}$$

Die Entscheidung wird also so getroffen:

$$\underbrace{c_1, \ldots, c_{n-1}}_{\text{ja}}, \underbrace{c_n}_{\text{würfeln}}, \underbrace{c_{n+1}, \ldots, c_{n'-1}}_{\text{unentsch.}}, \underbrace{c_{n'}}_{\text{würfeln}}, \underbrace{c_{n'+1}, \ldots, c_N}_{\text{nein}}.$$

Ein passender Score, der diese Treppenfunktion definiert, ist $T(c_i) = -i$.

Falls $n = n'$ und $p + q \leq 1$, sind μ und λ ebenfalls kompatibel, und wir können das Entscheidungsverfahren

$$\tau_{\lambda,\mu}(c) := \begin{cases} (1,0,0) & \text{für } 1 \leq i < n, \\ (p, 1-p-q, q) & \text{für } i = n, \\ (0,0,1) & \text{für } n < i \leq N. \end{cases}$$

definieren.

Hauptsatz 1 (Optimierung der stochastischen Entscheidung) *Das stochastische Entscheidungsverfahren $\tau_{\lambda,\mu}$ hat den Homonymfehler μ und den Synonymfehler λ; es hat unter allen stochastischen Entscheidungsverfahren, deren Fehlerraten durch μ und λ beschränkt sind, die kleinste Unentschiedenheitsrate.*

Beweis. Siehe [3]. $\diamond$

1.6 Die Gewichtsfunktion

In der Praxis ist man meist nicht in der Lage, die optimale Entscheidungsfunktion $\tau_{\lambda,\mu}$ exakt bestimmen zu können. Man müsste dazu ja für alle Beobachtungsdatensätze $c \in \Gamma$ die Wahrscheinlichkeiten $m(c) = P(c|M)$ und $u(c) = P(c|U)$ kennen. Auch die Herstellung der linearen Ordnung $\prec$ auf $\tilde{\Gamma}$ ist nicht unbedingt durchführbar. Daher sind Vereinfachungen wünschenswert.

Betrachtet man anstatt der Nummerierung auf $\tilde{\Gamma}$ als Score die Funktion $\frac{m}{u}$ selbst, erhält man eine Entscheidungsfunktion $\tilde{\tau}_{\lambda,\mu}$, die sich nur geringfügig von $\tau_{\lambda,\mu}$ unterscheidet:

- Ist $\frac{m}{u}(c) > \frac{m}{u}(c_n)$, so ist $c \prec c_n$, also $\tilde{\tau}_{\lambda,\mu}(c) = \tau_{\lambda,\mu}(c) = (1,0,0)$.
- Ist $\frac{m}{u}(c) = \frac{m}{u}(c_n)$ mit $c \neq c_n$, so kommt es auf die willkürliche Nummerierung an, ob $c \prec c_n$ oder $c_n \prec c$. Hier ist also der randomisierte Wert $\tilde{\tau}_{\lambda,\mu}(c) \neq \tau_{\lambda,\mu}(c) = (1,0,0)$ oder $(0,1,0)$.
- Analoges gilt an der Grenze zwischen unentschieden und nein.

Sind die Werte $\frac{m}{u}(c_n)$ und $\frac{m}{u}(c_{n'})$ also nicht einzig, ist $\tilde{\tau}_{\lambda,\mu}$ unter Umständen ein klein wenig von der Optimalität entfernt. [**Vermutung:** Auch dieses Entscheidungsverfahren ist optimal.]

Auf $\frac{m}{u}$ kann man ohne weiteren Verlust noch eine streng monotone Funktion $f\colon [0,\infty] \longrightarrow [0,\infty]$ anwenden und aufgrund des transformierten Scores $f \circ \frac{m}{u}$ entscheiden (auch wenn das mathematisch irrelevant ist). Üblich ist

$f = \log$, wobei es natürlich egal ist, zu welcher Basis man den Logarithmus bildet. In den Beispielen wird die Basis 10 verwendet. Die Funktion

$$w := \log \circ \frac{m}{u} : \tilde{\Gamma} \longrightarrow [-\infty, \infty]$$

heißt **Gewichtsfunktion** zur Beobachtungsfunktion γ; für $c \in \tilde{\Gamma}$ heißt $w(c)$ das **Gewicht** der Beobachtung c. Ist c auf M wahrscheinlicher als auf U, so ist $w(c) > 0$, ist c dagegen auf U wahrscheinlicher, so $w(c) < 0$.

> [Für $c \in \Gamma - \tilde{\Gamma}$ kann man $w(c)$ beliebig festlegen, etwa $= 0$; der Fall wird sowieso nie beobachtet.]

Das fast optimale Entscheidungsverfahren $\tilde{\tau}_{\lambda,\mu}$ ist dann das Schwellenwert-Verfahren zur Gewichtsfunktion w mit den Schwellenwerten

$$t_0 = w(c_n), \quad t_1 = w(c_{n'})$$

und geeigneter Randomisierung in den Grenzfällen.

1.7 Gruppen unabhängiger Merkmale

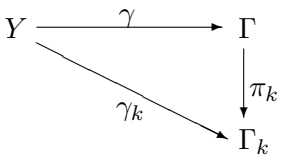
Eine weitere Vereinfachung betrifft die Zerlegung des Merkmalsraums in stochastisch unabhängige Komponenten:

$$\Gamma = \Gamma_1 \times \cdots \times \Gamma_K,$$

wobei die Γ_k selbst noch jeweils mehrere Merkmale zusammenfassen können.

> Typische unabhängige Merkmale sind Name und Geburtsjahr oder Geburtsmonat und Geburtsjahr (wenn man die kleine Ungenauigkeit durch die Schaltjahre außer Acht lässt).
>
> Typische Abhängigkeiten bestehen z. B. zwischen Namen und phonetischem Code des Namens – diese beiden Merkmale müsste man in einer Gruppe beieinander lassen. Abhängigkeit besteht auch zwischen Geschlecht und Vorname sowie, in abgemilderter Form, zwischen Geburtsjahr und Vorname.

Die Projektionen auf die Komponenten werden mit $\pi_k : \Gamma \longrightarrow \Gamma_k$ bezeichnet, die Komponenten der Beobachtungsfunktion entsprechend mit $\gamma_k := \pi_k \circ \gamma$.



Die Wahrscheinlichkeiten verhalten sich dann multiplikativ:

$$\begin{aligned} m(c_1,\ldots,c_K) &= P(c_1|M)\cdots P(c_K|M) &= m_1(c_1)\cdots m_K(c_K), \\ u(c_1,\ldots,c_K) &= P(c_1|U)\cdots P(c_K|U) &= u_1(c_1)\cdots u_K(c_K), \end{aligned}$$

ebenso der Quotient

$$\frac{m}{u}(c) = \frac{m_1}{u_1}(c_1)\cdots\frac{m_K}{u_K}(c_K).$$

Die Gewichtsfuktion als Logarithmus davon lässt sich schließlich additiv zerlegen:

$$w(c) = w_1(c_1) + \cdots + w_K(c_K)$$

mit

$$w_k(c_k) = \log\frac{m_k(c_k)}{u_k(c_k)}.$$

2 Stochastisches Matchen

Das Ergebnis von Abschnitt 1 wird jetzt auf das Match-Problem angewendet. Da die zur Bestimmung der Gewichte benötigten Größen teilweise unbekannt sind, werden auch Formeln angegeben, in denen diese adäquat durch Näherungswerte ersetzt werden.

2.1 Datenerzeugende Prozesse und ihre Fehler

Zugrunde liegt eine Menge Z (eine gedachte Population). Die Elemente von Z werden durch je einen Datensatz

$$\xi\colon Z \longrightarrow X$$

beschrieben, also durch eine Abbildung ξ in eine Menge $X = X_1 \times \cdots \times X_K$ von „Merkmalskombinationen“; jede der Mengen $X_1, \ldots, X_K$ ist endlich (ein „Merkmalsraum“). Dabei muss ξ nicht injektiv sein, d. h., „echte Homonyme“ sind nicht ausgeschlossen. Wir stellen uns vor, dass ξ die wahren Werte der Elemente von Z beschreibt.

Auf Z (oder einer Teilmenge davon) sei ein „**datenerzeugender Prozess**“ gegeben; das ist eine möglicherweise fehlerbehaftete Abbildung

$$\alpha\colon Z \longrightarrow X$$

in die Menge X; jeder der Merkmalsräume $X_1, \ldots, X_K$ kann auch ein Merkmal enthalten, das „fehlender Wert“ bedeutet.

Die Fehlerhaftigkeit von α wird beschrieben durch die **(werteabhängige) Fehlerrate**

$$\varepsilon_\alpha\colon X \longrightarrow [0,1], \qquad \begin{aligned}\varepsilon_\alpha(x) &= \frac{\#[\xi^{-1}x \cap (Z - \alpha^{-1}x)]}{\#\xi^{-1}x} \\ &= P(\alpha(z) \neq x \mid \xi(z) = x),\end{aligned}$$

sowie durch die **globale Fehlerrate**

$$\bar{\varepsilon}_\alpha = P(\alpha(z) \neq \xi(z)).$$

> Unter den Fehlerraten stellt man sich vor allem Eingabefehler bei der Erfassung der Daten vor. Zu beachten ist allerdings, dass die Wahrscheinlichkeit, dass sich ein Merkmal geändert hat, auch hierunter subsumiert ist – z. B. gibt es ja durchaus Gründe für einen Wechsel des Mermals „Name“ oder „Wohnort“.

Setzt man außerdem

$$p(x) = \frac{\#\xi^{-1}x}{\#Z}$$

als die relative Häufigkeit (oder Wahrscheinlichkeit), mit der der Wert $x \in X$ angenommen wird, so ist unmittelbar klar:

Hilfssatz 1 *Die globale Fehlerrate ist das gewichtete Mittel*

$$\bar{\varepsilon}_\alpha = \sum_{x \in X} p(x) \cdot \varepsilon_\alpha(x).$$

der werteabhängigen Fehlerraten.

In vielen Anwendungsfällen wird die Fehlerrate ε_α als unabhängig vom Wert x angenommen, also als konstant $= \bar{\varepsilon}_\alpha$.

2.2 Das Match-Problem

In der Grundmenge Z werden zwei Teilmengen $A, B \subseteq Z$ betrachtet, die sich möglicherweise überschneiden, und dazu die Mengen

$$\begin{aligned} M &:= \{(a,b) \in A \times B \mid a = b\} \quad \text{der „Matches“,} \\ U &:= \{(a,b) \in A \times B \mid a \neq b\} \quad \text{der „unpassenden Paare“.} \end{aligned}$$

Es ist $A \times B = M \,\dot\cup\, U$ disjunkte Vereinigung.

> **Vorstellung:** Die passenden Paare („Matches“) sollen gefunden werden, aber unter dem Handicap, dass die Elemente von A und B unvollständig und ungenau beschrieben sind. Diese Beschreibungen werden durch datenerzeugende Prozesse wie in Abschnitt 2.1 modelliert.

Auf den Teilmengen $A, B \subseteq Z$ sei jeweils ein datenerzeugender Prozess

$$\alpha \colon A \longrightarrow X, \quad \beta \colon B \longrightarrow X$$

gegeben; dazu gehören die **Fehlerraten** ε_α und ε_β. Diese beiden Abbildungen werden mit einer „Vergleichsfunktion“ δ zu einer Beobachtungsfunktion

$$\gamma \colon A \times B \xrightarrow{(\alpha,\beta)} X \times X \xrightarrow{\delta} \Gamma$$

zusammengefasst.

> **Vorstellung:** Die Bilder $\alpha(A)$ und $\beta(B)$ sind Dateien oder Datenbanken, in denen ein Teil der Grundpopulation Z abgebildet ist. Diese beiden Dateien sollen vereinigt werden. Aufgrund der „Vergleichsdatensätze“ $\gamma(a,b)$ soll mit möglichst geringem Fehler für jedes Paar $a \in A$ und $b \in B$ entschieden werden, ob $(a,b) \in M$ oder $(a,b) \in U$, d. h., ob $a = b$ oder $a \neq b$. Die Vergleichsfunktion δ nimmt im einfachsten Fall auf einer Komponente von X nur die beiden Werte „gleich“ oder „ungleich“ an, oft sind aber auch kompliziertere Vergleichsergebnisse notwendig; Beispiele folgen.

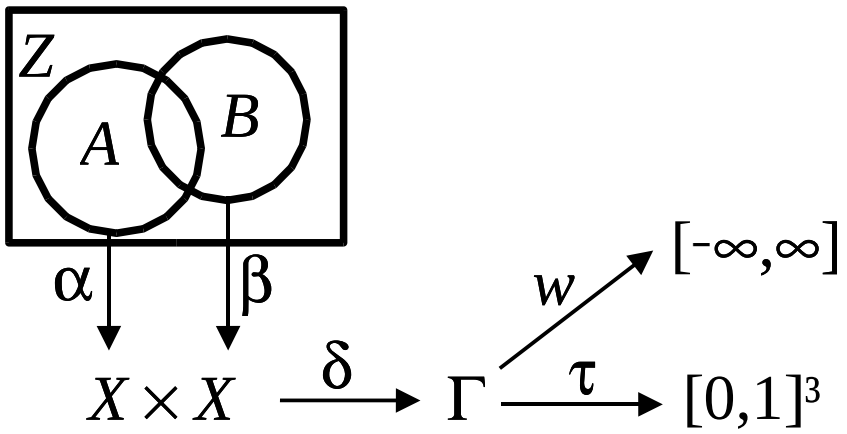


Abbildung 4: Die Match-Situation

> In der praktischen Anwendung enthalten die abzugleichenden Dateien oft noch weitere (unterschiedliche) Merkmale. Für den Match-Prozess werden aber nur gemeinsame Merkmale der beiden Datenquellen betrachtet; daher ist es in diesem Zusammenhang keine Einschränkung der Allgemeinheit, für die datenerzeugenden Prozesse α und β sowie für die „vollständige Beschreibung“ ξ den gleichen Satz von Merkmalen anzunehmen.

Damit liegt eine Entscheidungssituation wie in Kapitel 1 vor. Ein **stochastisches Matchverfahren** ist eine Entscheidungsfunktion

$$\tau : \Gamma \longrightarrow [0,1]^3.$$

Der Fehler erster Art heißt bei einem Match-Verfahren **Homonymfehler**, der Fehler zweiter Art **Synonymfehler**. Das optimale Entscheidungsverfahren aus Abschnitt 1.5 ergibt jetzt auch ein optimales Match-Verfahren; es minimiert bei gegebener Vergleichsfunktion und vorgegebenen Schranken für Homonym- und Synonymfehler die Zahl der unentschiedenen Fälle.

Dabei sind allerdings in der praktischen Anwendung einige der Größen unbekannt und müssen durch bekannte Größen geschätzt werden. Dazu versucht man, den Beobachtungsraum Γ in unabhängige Komponenten zu zerlegen, auf denen man den entsprechenden Summanden der Gewichtsfunktion bestimmen oder wenigstens hinreichend gut schätzen kann. Für jede Komponente des Beobachtungsraums ist das entsprechende Gewicht $w(c)$ der Logarithmus des Quotienten

$$\frac{P(c|M)}{P(c|U)};$$

im Zähler steht die Wahrscheinlichkeit, dass c bei passenden Paaren, im Nenner die, dass c bei unpassenden Paaren beobachtet wird. Diese Größen sind eigentlich erst *nach* dem Match-Vorgang oder zumindest nach einem Vorlauf mit Stichproben aus der gleichen Population gewinnbar, und auch das nur mit einigem Aufwand. Es gibt aber brauchbare Schätzungen, die mit wenig Aufwand auszuwerten sind.

Annahmen, die realistisch sind und im folgenden stets getroffen werden, sind:

- Die Ausprägung $x \in X$ kommt in Z sowie in A und B und auch in $A \cap B$ mit der Wahrscheinlichkeit

$$\begin{aligned} p(x) &= P\{z \in Z \mid \xi(z) = x\} \\ &\approx P\{a \in A \mid \xi(a) = x\} \approx P\{b \in B \mid \xi(b) = x\} \\ &\approx P\{a \in A \cap B \mid \xi(a) = x\} \end{aligned}$$

  vor.

- Die Fehlerraten in den datenerzeugenden Prozessen werden auf A und $A \cap B$ durch ε_α, auf B und $A \cap B$ durch ε_β hinreichend genau geschätzt.

- Die Fehlerraten in den datenerzeugenden Prozessen sind klein oder heben sich in etwa auf, so dass

$$P(\alpha(a) = x) \approx P(\beta(b) = x) \approx p(x)$$

  gilt.

- $\#M$ ist klein gegen $\#A \cdot \#B$.

2.3 Globale Gewichte

Ohne wesentliche Einschränkung wird im Rest dieses Kapitels angenommen, dass der Merkmalsraum X nur eine Komponente enthält. Die Verteilung der Ausprägungen sei bekannt. Der Vergleichsraum Γ enthalte nur die beiden Elemente „gleich“ und „ungleich“, und die Vergleichsfunktion $\delta : X \times X \longrightarrow \Gamma$ sei sinngemäß definiert. In diesem Fall spricht man nach Newcombe [7] von **globalen Gewichten**, da die spezifischen Werte der Merkmale beim Match-Verfahren gar nicht berücksichtigt werden.

Die Funktion $u : \Gamma \longrightarrow [0, 1]$, die die bedingte Wahrscheinlichkeit $P(\text{„(un)gleich“}|U)$ für einen Vergleichswert bei *nicht zusammengehörigen*

Paaren angibt, hat dann die folgenden Werte:

$$\begin{aligned} u(\text{„gleich“}) &= P(\alpha(a) = \beta(b) \mid a \in A, b \in B, a \neq b) \\ &= \sum_{x \in X} P(a \neq b, \alpha(a) = x, \beta(b) = x) \\ &\approx \sum_{x \in X} P(\alpha(a) = x, \beta(b) = x) \\ &\approx \sum_{x \in X} p(x)^2 \end{aligned}$$

Da es nur zwei mögliche Vergleichswerte gibt, ist

$$u(\text{„ungleich“}) \approx 1 - \sum_{x \in X} p(x)^2.$$

Besonders einfach werden diese Formeln für die Gleichverteilung: Hat X n Ausprägungen x mit jeweils der Wahrscheinlichkeit $p(x) = \frac{1}{n}$, so ist

$$\begin{aligned} u(\text{„gleich“}) &\approx \frac{1}{n}, \\ u(\text{„ungleich“}) &\approx 1 - \frac{1}{n}. \end{aligned}$$

Die Funktion $m : \Gamma \longrightarrow [0, 1]$, die die bedingte Wahrscheinlichkeit $P(\text{„(un)gleich“}|M)$ für einen Vergleichswert bei *zusammenpassenden* Paaren angibt, hat die Werte:

$$\begin{aligned} m(\text{„ungleich“}) &= P(\alpha(a) \neq \beta(b) \mid a \in A \cap B) \\ &= P(\alpha(a) \neq \xi(a), \beta(a) = \xi(a) \mid a \in A \cap B) \\ &\quad + P(\alpha(a) = \xi(a), \beta(a) \neq \xi(a) \mid a \in A \cap B) \\ &\quad + P(\alpha(a) \neq \xi(a), \beta(a) \neq \xi(a), \alpha(a) \neq \beta(a) \mid a \in A \cap B) \end{aligned}$$

$$\approx \sum_{x \in X} p(x) \cdot \left[\varepsilon_\alpha(x)\,(1 - \varepsilon_\beta(x)) + \varepsilon_\beta(x)\,(1 - \varepsilon_\alpha(x)) + \varepsilon_\alpha(x)\varepsilon_\beta(x)\right],$$

wobei die Wahrscheinlichkeit, dass α und β zufällig den gleichen falschen Wert ergeben haben, vernachlässigt wird. Ersetzt man in dem Restterm, der klein von zweiter Ordnung ist, noch $\varepsilon_\beta(x)$ durch die globale Fehlerrate $\bar{\varepsilon}_\beta$, setzt sich die Formel fort zu

$$\begin{aligned} m(\text{„ungleich“}) &\approx \bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta - \sum_{x \in X} p(x) \cdot \varepsilon_\alpha(x)\varepsilon_\beta(x) \\ &\approx \bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta - \bar{\varepsilon}_\alpha \cdot \bar{\varepsilon}_\beta, \end{aligned}$$

und dazu passend

$$m(\text{„gleich“}) \approx 1 - \bar{\varepsilon}_\alpha - \bar{\varepsilon}_\beta + \bar{\varepsilon}_\alpha \cdot \bar{\varepsilon}_\beta = (1 - \bar{\varepsilon}_\alpha) \cdot (1 - \bar{\varepsilon}_\beta).$$

Damit ist gezeigt:

Satz 1 *Wird im Vergleichsraum nur zwischen „gleich“ und „ungleich“ unterschieden und sind die globalen Fehlerraten $\bar{\varepsilon}_\alpha$ und $\bar{\varepsilon}_\beta$ klein, so gelten die Näherungsformeln*

$$\begin{aligned} w(\text{„}gleich\text{“}) &\approx \log \frac{(1-\bar{\varepsilon}_\alpha)\cdot(1-\bar{\varepsilon}_\beta)}{\sum_{x\in X} p(x)^2}, \\ w(\text{„}ungleich\text{“}) &\approx \log \frac{\bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta - \bar{\varepsilon}_\alpha\cdot\bar{\varepsilon}_\beta}{1-\sum_{x\in X} p(x)^2}. \end{aligned}$$

Korollar 1 *Der Merkmalsraum X sei gleichverteilt mit n Ausprägungen. Die globalen Fehlerraten $\bar{\varepsilon}_\alpha$ und $\bar{\varepsilon}_\beta$ seien klein. Dann ist*

$$\begin{aligned} w(\text{„}gleich\text{“}) &\approx \log[n\cdot(1-\bar{\varepsilon}_\alpha)(1-\bar{\varepsilon}_\beta)], \\ w(\text{„}ungleich\text{“}) &\approx \log \frac{n(\bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta - \bar{\varepsilon}_\alpha\cdot\bar{\varepsilon}_\beta)}{n-1}. \end{aligned}$$

Das Gewicht bei Gleichheit wächst also im wesentlichen proportional zu $\log n$, während das Gewicht bei Ungleichheit praktisch von n unabhängig ist. Fehlerraten um die 10% sind für Satz 1 samt Korollar 1 gut genug; siehe Abschnitt 2.4. Bei wesentlich größeren Fehlerraten wird jedes Match-Verfahren aber sowieso ziemlich sinnlos.

Aus dem Satz folgt auch, was eigentlich von vorneherein klar ist:

Korollar 2 *Sind die Eingabefehler* 0*, d. h., ist das Merkmal in X korrekt erfasst, so ist die Ungleichheit ein KO-Kriterium: Zwei Datensätze die in X nicht übereinstimmen, können nicht zusammengehören; die Ungleichheit hat das Gewicht $-\infty$.*

Beispiel 1. Das Merkmal sei das Geschlecht, also $n = 2$ bei Gleichverteilung. Die globale Fehlerrate sei in beiden Fällen $\varepsilon = \frac{1}{20} = 5\%$. Dann ist

$$\begin{aligned} w(\text{„gleich“}) &\approx \log\left(2\cdot\left(\frac{19}{20}\right)^2\right) \approx \log 1.8 \approx 0.26, \\ w(\text{„ungleich“}) &\approx \log\left(2\cdot\left(\frac{1}{10}-\frac{1}{400}\right)\right) \approx \log 0.195 \approx -0.71. \end{aligned}$$

Die Übereinstimmung des Geschlechts ergibt nur ein sehr kleines positives Gewicht für das Matchen, die Nichtübereinstimmung ein etwas größeres negatives; hier ist die Fehlerrate dafür ausschlaggebend, dass letzteres nicht größer ist.

Beispiel 2. Das Merkmal sei der Geburtsmonat, also $n = 12$. Die globale Fehlerrate sei ebenfalls $\varepsilon = 5\%$. Dann ist (wenn Gleichverteilung

angenommen wird)

$$\begin{aligned} w(\text{„gleich“}) &\approx \log\left(12 \cdot \left(\frac{19}{20}\right)^2\right) \approx \log 10.8 \approx 1.03, \\ w(\text{„ungleich“}) &\approx \log\left(\frac{12}{11} \cdot \left(\frac{1}{10} - \frac{1}{400}\right)\right) \approx \log 0.106 \approx -0.97. \end{aligned}$$

Die Übereinstimmung im Geburtsmonat hat natürlich ein größeres Gewicht als die Übereinstimmung im Geschlecht.

Beispiel 3. Lohnt es sich, die nicht ganz exakte Gleichverteilung der Monate zu berücksichtigen? Nehmen wir also

$$p(x) = \begin{cases} \frac{31}{365.25} & \text{für } x = \text{Januar, März, } \ldots, \\ \frac{30}{365.25} & \text{für } x = \text{April, Juni, } \ldots, \\ \frac{28.25}{365.25} & \text{für } x = \text{Februar.} \end{cases}$$

Die Zähler der Gewichte sind dann nach Satz 1

$$0.95 \cdot 0.95 = 0.9025 \quad \text{bzw.} \quad 1 - 0.9025 = 0.0975,$$

die Nenner

$$\sum p(x)^2 = 7 \cdot \left(\frac{31}{365.25}\right)^2 + 4 \cdot \left(\frac{30}{365.25}\right)^2 + \left(\frac{28.25}{365.25}\right)^2 \approx 0.0834$$

$$\text{bzw.} \quad 1 - 0.0834 = 0.9166.$$

Daraus berechnen sich die Gewichte zu

$$\begin{aligned} w(\text{„gleich“}) &\approx \log \frac{0.9025}{0.0835} = \log 10.77 \approx 1.03, \\ w(\text{„ungleich“}) &\approx \log \frac{0.0975}{0.9166} \approx \log 0.1064 \approx -0.97. \end{aligned}$$

Das legt den Schluss nahe: *Die Abweichung der Monatslängen von der Gleichverteilung braucht bei den globalen Gewichten nicht berücksichtigt zu werden.*

Beispiel 4. Wie wirken sich Änderungen der Fehlerrate aus? Nehmen wir an, die aus A mit α gewonnenen Daten haben nur die globale Fehlerrate $\bar{\varepsilon}_\alpha = 1\%$; $\bar{\varepsilon}_\beta$ bleibe bei 5%. Die Zähler der Gewichte sind dann nach Satz 1

$$0.99 \cdot 0.95 = 0.9405 \quad \text{bzw.} \quad 1 - 0.9405 = 0.0595,$$

die Nenner

$$\frac{1}{12} \quad \text{bzw.} \quad \frac{11}{12}.$$

Daraus berechnen sich die Gewichte zu

$$\begin{aligned} w(\text{„gleich“}) &\approx \log(12 \cdot 0.9405) = \log 11.3 \approx 1.05, \\ w(\text{„ungleich“}) &\approx \log(\frac{12}{11} \cdot 0.0595) \approx \log 0.0649 \approx -1.19. \end{aligned}$$

Das legt den Schluss nahe: *Eine Variation in den Fehlerraten braucht beim globalen Gewicht der Gleichheit praktisch nicht berücksichtigt zu werden, wirkt sich aber deutlich auf das Gewicht der Ungleichheit aus.*

2.4 Approximation globaler Gewichte nach Newcombe

Eine aus umfangreicher empirischer Erfahrung gewonnene Faustregel von Newcombe [7] sagt:

> Ein Faktor bis zu 2 kann bei den Wahrscheinlichkeitsquotienten („Likelihood Ratios“ oder „Odds“) für die einzelnen Merkmale vernachlässigt werden, ohne das Match-Verfahren nennenswert zu beeinträchtigen. Für die Gewichte bedeutet das (bei Logarithmen zur Basis 10) einen Summanden bis zu etwa 0.3.

Etwas konservativer sollen hier Gewichtssummanden bis 0.1, also Faktoren bis 1.25 vernachlässigt werden. Damit lässt sich Satz 1 samt dem ersten Korollar etwas vereinfachen; bei Fehlerraten um 5% oder kleiner sind die folgenden Näherungsformeln sicherlich gut genug.

Korollar 3 (Newcombe-Approximation der globalen Gewichte) *Wird im Vergleichsraum nur zwischen „gleich“ und „ungleich“ unterschieden und sind die globalen Fehlerraten $\bar{\varepsilon}_\alpha$ und $\bar{\varepsilon}_\beta$ klein, so gelten die Näherungsformeln*

$$\begin{aligned} w(\text{„}\textit{gleich}\text{“}) &\approx \log \frac{1}{\sum_{x \in X} p(x)^2}, \\ w(\text{„}\textit{ungleich}\text{“}) &\approx \log \frac{\bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta}{1 - \sum_{x \in X} p(x)^2}. \end{aligned}$$

Korollar 4 *Der Merkmalsraum X sei gleichverteilt mit n Ausprägungen. Die globalen Fehlerraten $\bar{\varepsilon}_\alpha$ und $\bar{\varepsilon}_\beta$ seien klein. Dann ist*

$$\begin{aligned} w(\text{„}\textit{gleich}\text{“}) &\approx \log n, \\ w(\text{„}\textit{ungleich}\text{“}) &\approx \log \frac{n(\bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta)}{n-1}. \end{aligned}$$

Ist n nicht ganz klein, vereinfacht sich die zweite Formel zu

$$w(\text{„}\textit{ungleich}\text{“}) \approx \log(\bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta).$$

Die letze Approximation kann nach der obigen Leitlinie für etwa $n \geq 6$ ohne Schaden verwendet werden.

Beispiel 1. Die vereinfachte Formel ergibt für die – als gleichverteilt angenommenen – Geburtsmonate mit beiden globalen Fehlerraten = 5% die Gewichte

$$\begin{aligned} w(\text{„gleich“}) &\approx \log 12 \approx 1.08, \\ w(\text{„ungleich“}) &\approx \log(0.10) \approx -1.00, \end{aligned}$$

die zwar etwas von den Werten aus Abschnitt 2.3 abweichen, aber weit unterhalb der durch die Faustregel von NEWCOMBE oder die konservative Leitlinie gegebenen Toleranz. NEWCOMBE selbst gibt in [7] folgende empirisch gewonnenen Werte für den Geburtsmonat an:

$$\begin{aligned} w_{\text{empirisch}}(\text{„gleich“}) &\approx 1.05, \\ w_{\text{empirisch}}(\text{„ungleich“}) &\approx -1.14, \end{aligned}$$

deren Abweichung zu den durch theoretische Überlegung gewonnenen Werten ebenfalls noch innerhalb der Toleranz liegt; umgekehrt kann man natürlich aus den empirischen Werten die Summe der globalen Fehlerraten in den verwendeten Daten bestimmen:

$$\bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta \approx 10^{-1.14} \approx 0.072 = 7.2\%.$$

Beispiel 2. Andere empirisch gewonnene globale Gewichte nach NEWCOMBE sind [7, S. 15f.]

$$\begin{aligned} w_{\text{empirisch}}(\text{„Name gleich“}) &\approx 2.98, \\ w_{\text{empirisch}}(\text{„Name ungleich“}) &\approx -1.46, \\ w_{\text{empirisch}}(\text{„Vorname gleich“}) &\approx 1.94, \\ w_{\text{empirisch}}(\text{„Vorname ungleich“}) &\approx -0.67, \\ w_{\text{empirisch}}(\text{„Geburtsjahr gleich“}) &\approx 1.85, \\ w_{\text{empirisch}}(\text{„Geburtsjahr ungleich“}) &\approx -0.64, \\ w_{\text{empirisch}}(\text{„Geburtstag gleich“}) &\approx 1.41, \\ w_{\text{empirisch}}(\text{„Geburtstag ungleich“}) &\approx -0.81; \end{aligned}$$

diese geben zumindest eine gute Vorstellung davon, welche Größenordnungen von Gewichten man bei verschiedenen Merkmalen erwarten kann. Die Gewichte für die Gleichheit hängen natürlich von der untersuchten Population, die für die Ungleichheit von den globalen Fehlerraten ab.

2.5 Werteabhängige Gewichte

Es soll jetzt nicht nur auf Gleichheit geprüft werden, sondern auch berücksichtigt werden, um welchen Wert es sich handelt. Dazu nimmt man als Vergleichsraum

$$\Gamma = X \times X$$

und als Vergleichsfunktion δ die identische Abbildung von $X \times X$. Für $x, x' \in X$ gilt dann:

$$\begin{aligned}
\gamma^{-1}(x,x') &= \{(a,b) \in A \times B \mid \alpha(a) = x, \beta(b) = x'\} \\
&= \alpha^{-1}(x) \times \beta^{-1}(x'), \\
M \cap \gamma^{-1}(x,x') &= \{(a,a) \mid a \in A \cap B, \alpha(a) = x, \beta(a) = x'\} \\
U \cap \gamma^{-1}(x,x') &= \{(a,b) \in A \times B \mid a \neq b, \alpha(a) = x, \beta(b) = x'\}.
\end{aligned}$$

Damit gilt

$$u(x,x') = \frac{\#(U \cap \gamma^{-1}(x,x'))}{\#U} \approx \frac{\#(\alpha^{-1}(x) \times \beta^{-1}(x'))}{\#(A \times B)} = p(x)p(x')$$

für alle $x, x' \in X$, da die Bedingung $a \neq b$ hier vernachlässigt werden kann.

Etwas komplizierter ist

$$m(x,x') = \frac{\#(M \cap \gamma^{-1}(x,x'))}{\#M} = \frac{\#\{a \in A \cap B \mid \alpha(a) = x, \beta(a) = x'\}}{\#A \cap B}$$

zu schätzen. Im 1. *Fall* sei $x = x'$. Dann ist

$$m(x,x) = P(\alpha(a) = x \mid a \in A \cap B) \cdot P(\beta(a) = x \mid a \in A \cap B, \alpha(a) = x).$$

Für die Fälle von a, wo x der „wahre“ Wert $\xi(a)$ ist, ergibt das die Faktoren $p(x)(1-\varepsilon_\alpha(x))$ für $\alpha(a) = x$ und $1-\varepsilon_\beta(x)$ für $\beta(a) = x$ zusätzlich. Zusammen macht das unter den Annahmen aus Abschnitt 2.2:

$$m(x,x) \approx p(x) \cdot (1 - \varepsilon_\alpha(x))(1 - \varepsilon_\beta(x)).$$

Im 2. *Fall* sei $x \neq x'$. Zunächst ein Hilfssatz:

Hilfssatz 2 *Unter der Annahme, dass bei α jeder falsche Wert mit gleicher Wahrscheinlichkeit angenommen wird, gilt für $x \neq x'$ auf A und allen genügend großen Teilmengen davon:*

$$P(\xi(a) = x, \alpha(a) = x') = \frac{p(x)\varepsilon_\alpha(x)}{n-1},$$

wobei $n = \#X$.

Beweis. Sei η der von x' unabhängige Wert $P(\xi(a) = x, \alpha(a) = x')$. Dann gilt

$$\begin{aligned} p(x) &= P(\xi(a) = x) \\ &= \sum_{y \in X} P(\xi(a) = x, \alpha(a) = y) \\ &= P(\xi(a) = x, \alpha(a) = x) + \sum_{y \in X - \{x\}} P(\xi(a) = x, \alpha(a) = y) \\ &= p(x) \cdot (1 - \varepsilon_\alpha(x)) + (n-1) \cdot \eta. \end{aligned}$$

Daraus folgt die Behauptung. $\diamond$

Für β wird ebenfalls die Annahme des Hilfssatzes gemacht. Dann tragen zu

$$m(x, x') = P(\alpha(a) = x, \beta(a) = x')$$

auf $A \cap B$ die Fälle a, wo $x = \xi(a)$ der „wahre“ Wert ist, den Summanden

$$p(x) \cdot (1 - \varepsilon_\alpha(x)) \cdot \frac{\varepsilon_\beta(x')}{n-1}$$

bei, die Fälle, wo $x' = \xi(a)$, den Summanden

$$p(x') \cdot (1 - \varepsilon_\beta(x')) \cdot \frac{\varepsilon_\alpha(x)}{n-1},$$

die Fälle wo x und x' falsche Werte sind ($y = \xi(a)$ der richtige), den Summanden

$$\approx \frac{p(y)\varepsilon_\alpha(y)}{n-1} \cdot \frac{p(y)\varepsilon_\beta(y)}{n-1},$$

der vernachlässigt werden kann. Das ergibt die Näherung

$$m(x, x') \approx \frac{1}{n-1} \left[p(x)\varepsilon_\beta(x') + p(x')\varepsilon_\alpha(x) - (p(x) + p(x')) \cdot \varepsilon_\alpha(x)\varepsilon_\beta(x') \right],$$

oder, wenn das Produkt $\varepsilon_\alpha(x)\varepsilon_\beta(x')$ gegenüber $\varepsilon_\alpha(x)$ und $\varepsilon_\beta(x')$ vernachlässigt wird,

$$m(x, x') \approx \frac{p(x) \cdot \varepsilon_\beta(x') + p(x') \cdot \varepsilon_\alpha(x)}{n-1}.$$

Satz 2 *Ist der Vergleichsraum $= X \times X$ und sind die Eingabefehlerraten klein, so gelten die Näherungsformeln*

$$\begin{aligned} w(x, x) &\approx \log \frac{(1 - \varepsilon_\alpha(x))(1 - \varepsilon_\beta(x))}{p(x)} \\ w(x, x') &\approx \log \left[\frac{1}{n-1} \left(\frac{\varepsilon_\alpha(x)}{p(x)} + \frac{\varepsilon_\beta(x')}{p(x')} \right) \right] \quad \textit{für } x \neq x'. \end{aligned}$$

Auch hier dürfen die Eingabefehlerraten um 10% liegen; sind sie nur etwa 5% oder kleiner, kann man wieder vereinfachen:

Korollar 1 (NEWCOMBE-Approximation der werteabhängigen Gewichte) *Bei gleichen Voraussetzungen mit kleineren Fehlerraten gilt:*

$$\begin{aligned} w(x,x) &\approx -\log p(x), \\ w(x,x') &\approx \log\left[\frac{\varepsilon_\alpha(x)}{p(x)} + \frac{\varepsilon_\beta(x')}{p(x')}\right] - \log(n-1) \quad \textit{für } x \neq x'. \end{aligned}$$

Korollar 2 *Der Merkmalsraum X sei gleichverteilt mit n Ausprägungen. Die Fehlerraten ε_α und ε_β seien klein und unabhängig vom Wert. Dann ist*

$$\begin{aligned} w(x,x) &\approx \log[n \cdot (1-\bar{\varepsilon}_\alpha) \cdot (1-\bar{\varepsilon}_\beta)], \\ w(x,x') &\approx \log \frac{n(\bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta)}{n-1} \quad \textit{für } x \neq x'. \end{aligned}$$

D. h., bei Gleichverteilung ändert sich durch die Berücksichtigung der Werte beim Vergleich praktisch nichts am Gewicht gegenüber der bloßen Unterscheidung „gleich oder ungleich“ – *die werteabhängigen Gewichte sind bei gleichverteilten Merkmalen und werteunabhängigen Fehlerraten im wesentlichen gleich den globalen Gewichten.*

Beispiel 1. Lohnt es sich, die Ungleichverteilung der Monatslängen bei den Gewichten für den Geburtsmonat zu berücksichtigen? Für die Gleichheit berechnen sich die werteabhängigen Gewichte zu

$$w(x,x) = \begin{cases} -\log \frac{31}{365.25} = 1.07 & \text{für Januar, März, \ldots,} \\ -\log \frac{30}{365.25} = 1.09 & \text{für April, Juni, \ldots,} \\ -\log \frac{28.5}{365.25} = 1.11 & \text{für Februar,} \end{cases}$$

also etwas höhere Werte als das globale Gewicht, besonders für den etwas selteneren Monat Februar. Bei der Ungleichheit nehmen wir wieder konstante Fehlerraten von jeweils 5% an. Dann ergibt die Formel

$$\begin{aligned} w(x,x') &\approx \log(0.05 \cdot 365.25) - \log 11 + \log\left[\frac{1}{t_1} + \frac{1}{t_2}\right] \\ &\approx 0.220 + \log\left[\frac{1}{t_1} + \frac{1}{t_2}\right] \end{aligned}$$

für zwei verschiedene Monate mit t_1 bzw. t_2 Tagen. Daraus ergeben sich für die Ungleichheit die Gewichte

	31	30	28.5
31	-0.97	-0.96	-0.95
30	-0.96	-0.95	-0.94
28.5	-0.95	-0.94	–

also auch hier eine geringe Abweichung vom globalen Gewicht der Ungleichheit. Ob man diese Abweichungen zum Anlass nehmen sollte, für die Monate werteabhängige Gewichte zu verwenden, sei dahingestellt. Nach der Faustregel von NEWCOMBE bräuchte man es nicht, nach der konservativeren Richtlinie aus 2.4 auch nicht, obwohl man schon etwas näher an der Grenze ist.

Beispiel 2. Deutlicher ist der Effekt der Ungleichverteilung bei den Tagen $1, \ldots, 31$. Nehmen wir wieder konstante Fehlerraten von $\varepsilon = 5\%$ an, so sind die globalen Gewichte

$$\begin{aligned} w(\text{„gleich“}) &\approx \log 31 \approx 1.49, \\ w(\text{„ungleich“}) &\approx \log 0.1 = -1.00. \end{aligned}$$

Berücksichtigt man dagegen, dass die Tage $1, \ldots 28$ in einem Jahr mit insgesamt 365.25 Tagen je 12-mal vorkommen, der Tag 29 nur 11.25-mal, der Tag 30 nur 11-mal und der Tag 31 sogar nur 7-mal vorkommt, so sind die Gewichte für die Gleichheit

$$w(x,x) = \begin{cases} -\log \frac{12}{365.25} = 1.48 & \text{für } x = 1, \ldots, 28, \\ -\log \frac{11.25}{365.25} = 1.51 & \text{für } x = 29 \\ -\log \frac{11}{365.25} = 1.52 & \text{für } x = 30 \\ -\log \frac{7}{365.25} = 1.72 & \text{für } x = 31. \end{cases}$$

Für die Ungleichheit gilt die Formel

$$\begin{aligned} w(x,x') &\approx \log(0.05 \cdot 365.25) - \log 30 + \log \left[\frac{1}{t_1} + \frac{1}{t_2} \right] \\ &\approx -0.216 + \log \left[\frac{1}{t_1} + \frac{1}{t_2} \right] \end{aligned}$$

für zwei verschiedene Tage mit t_1 bzw. t_2 Vorkommen im Jahr. Daraus ergeben sich für die Ungleichheit die Gewichte

	≤ 28	29	30	31
≤ 28	-0.99	-0.98	-0.97	-0.86
29	-0.98	–	-0.96	-0.85
30	-0.97	-0.96	–	-0.85
31	-0.86	-0.85	-0.85	–

Die Abweichungen zu den globalen Gewichten sind also nach der Faustregel von NEWCOMBE gerade noch, nach der konservativeren Regel allerdings nicht zu vernachlässigen.

2.6 Beispiel: Merkmale mit zwei Ausprägungen

Der Merkmalsraum $X = \{x, x'\}$ sei zweielementig mit $p(x) = q$, $p(x') = 1 - q$. Dann ist

$$p(x)^2 + p(x')^2 = q^2 + (1-q)^2 = 1 - 2q + 2q^2.$$

Im folgenden werden die zwei verschiedenen Vergleichsfunktionen aus 2.3 und 2.5 betrachtet, also die globalen und die lokalen Gewichte bestimmt. In beiden Fällen werden die Fehlerraten $\varepsilon = \varepsilon_\alpha = \varepsilon_\beta$ als konstant und gleich angenommen.

Globale Gewichte

Wie in Abschnitt 2.3 wird der Vergleichsraum Γ zweielementig gewählt. Damit ist

$$\begin{aligned} w(\text{„gleich“}) &\approx \log \frac{1-2\varepsilon}{1-2q+2q^2}, \\ w(\text{„ungleich“}) &\approx \log \frac{2\varepsilon}{2q-2q^2} = \log \frac{\varepsilon}{q(1-q)}. \end{aligned}$$

Nimmt man $\varepsilon = \frac{1}{10} = 10\%$ an, so ergeben sich für $q = \frac{1}{2}$ die Werte

$$\begin{aligned} w(\text{„gleich“}) &\approx \log[2 \cdot (1-2\varepsilon)] = \log 1.6 \approx 0.20, \\ w(\text{„ungleich“}) &\approx \log 4\varepsilon = \log 0.4 \approx -0.40, \end{aligned}$$

für $q = \frac{1}{4}$ die Werte

$$\begin{aligned} w(\text{„gleich“}) &\approx \log \left[\frac{8 \cdot (1-2\varepsilon)}{5}\right] = \log 1.28 \approx 0.11, \\ w(\text{„ungleich“}) &\approx \log \frac{16\varepsilon}{3} = \log 0.53 \approx -0.27. \end{aligned}$$

Da die Fehlerraten mit 10% nicht besonders klein sind, wurde hier nicht die Newcombe-Approximation, sondern die etwas kompliziertere genauere Formal verwendet.

Werteabhängige Gewichte

Nach den Formeln in 2.5 gilt

$$\begin{aligned} w(x,x) &\approx \log \frac{1-2\varepsilon}{q}, \\ w(x',x') &\approx \log \frac{1-2\varepsilon}{1-q}, \\ w(x,x') = w(x',x) &\approx \log \frac{\varepsilon}{q(1-q)}. \end{aligned}$$

Speziell für $\varepsilon = \frac{1}{10}$ und $q = \frac{1}{4}$ wird daraus

$$\begin{aligned} w(x,x) &\approx \log\frac{32}{10} \approx 0.51, \\ w(x',x') &\approx \log\frac{32}{30} \approx 0.03, \\ w(x,x') = w(x',x) &\approx \log\frac{16}{30} \approx -0.27; \end{aligned}$$

hier ergibt sich bei Ungleichheit keine Änderung im Gewicht, wohl aber bei Gleichheit: Im „seltenen“ Fall ist das Gewicht der Übereinstimmung deutlich vergrößert, während die Gleichheit im „häufigen“ Fall geringer gewichtet wird.

2.7 Komplexe Merkmale

$\ldots (x,y)$ mit y abhängig von x $\ldots$

3 Algorithmus

Der zu implementierende Match-Vorgang sieht so aus („iterativer Abgleich“): Es gibt eine bestehende Datenbank B mit bisher erfassten Fällen. Kommt ein neuer Fall hinzu, wird nachgesehen, ob dieser in der Datenbank schon vorhanden ist, d. h., ob der Match-Algorithmus genau einen (plausiblen) Treffer liefert.

- Falls das so ist, wird der neue Fall nicht zusätzlich in die Datenbank aufgenommen; ist die Übereinstimmung mit dem vorhandenen Fall allerdings nicht perfekt, so liegen neue Eingabefehler vor, und die Fehlerraten sind geeignet anzupassen.
- Liegt kein Treffer vor, wird der neue Fall zusätzlich in die Datenbank aufgenommen; gegebenenfalls werden die Merkmalshäufigkeiten entsprechend angepasst.
- Gibt es mehr als einen Treffer oder unentscheidbare Fälle, wird der neue Fall als unentscheidbar zurückgewiesen.

Angenommen wird, dass die in Abschnitt 2 hergeleiteten Gewichtsformeln auch für diesen schrittweisen Ablauf des Match-Vorgangs geeignet sind.

3.1 Zusammenfassung der benötigten Formeln

Für das Match-Verfahren wird der Merkmalsraum in eine direkte Summe von stochastisch unabhängigen Komponenten zerlegt. Für jede solche Komponente X und für jedes zu vergleichende Wertepaar $x, x' \in X$ wird ein Gewicht berechnet; die Gewichte aller Komponenten werden danach addiert. Bei fehlenden Werten in einer Komponente wird das Gewicht für das jeweilige Merkmal = 0 gesetzt; das läuft darauf hinaus, dass das Merkmal für diesen Vergleich nicht berücksichtigt wird.

Für die Bestimmung des Gewichts einer Komponente durch eine Näherungsformel wird hier stets angenommen, dass die Fehlerraten ε_α und ε_β bei beiden datenerzeugenden Prozessen höchstens in der Größenordnung von 5% liegen. Die Wahrscheinlichkeit $p(x)$, mit der in der Merkmalskomponente X ein beliebiger Wert $x \in X$ angenommen wird, sei bekannt oder hinreichend genau ermittelbar.

Dann wird das Gewicht, je nach Übereinstimmung oder Nichtübereinstimmung nach den Näherungsformeln

$$\begin{aligned} w(x,x) &\approx -\log p(x), \\ w(x,x') &\approx \log\left[\frac{\varepsilon_\alpha(x)}{p(x)} + \frac{\varepsilon_\beta(x')}{p(x')}\right] - \log(n-1) \quad \text{für } x \neq x', \end{aligned}$$

berechnet, wobei $n = n_X = \#X$ die Anzahl der Elemente von X ist. Der Fall $n = 1$ kann hier ausgeschlossen werden, da ein konstantes Merkmal nicht zum Vergleich geeignet ist.

Ist X gleichverteilt, so braucht man nur die globalen Gewichte und kann diese nach den einfacheren Näherungsformeln

$$\begin{aligned} w(\text{„gleich“}) &\approx \log n, \\ w(\text{„ungleich“}) &\approx \log \frac{n\eta}{n-1} = \log\left(1 + \frac{1}{n-1}\right) + \log \eta, \end{aligned}$$

bestimmen, wobei $\eta = \bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta$ die Summe der globalen Fehlerraten ist; für $n \geq 6$ wird die letzte Formel zu

$$w(\text{„ungleich“}) \approx \log \eta,$$

vereinfacht. Dabei wird auch angenommen, dass bei gleichverteilten Merkmalen die Fehlerraten im wesentlichen vom Wert unabhängig sind.

Die Wahrscheinlichkeiten $p(x)$ können durch die jeweilige relative Häufigkeit geschätzt werden, mit der der Wert x bisher in der Datenbank B enthalten ist, also durch die Formel

$$p(x) \approx \frac{N_X(x)}{N_B},$$

wobei N_B die Gesamtzahl der aktuell in B vorhandenen Datensätze ist und $N_X(x)$ die Anzahl der Datensätze darunter, die im betrachteten Merkmal den Wert x haben. Diese Zahlen müssen also beim iterativen Abgleich stets mitgeführt werden.

Die Schätzung der Fehlerraten ist komplizierter und kommt nicht ohne zusätzliche Annahmen aus. In den beiden folgenden Abschnitten wird jeweils eine Schätzung unter einer solchen Zusatzannahme hergeleitet.

3.2 Schätzung der Fehlerraten im Modell I

In diesem vereinfachten Modell wird angenommen, dass *beide Fehlerraten gleich*, also durch eine gemeinsame Funktion $\varepsilon\colon A \cup B \longrightarrow [0,1]$ gegeben sind.

Globale Fehlerraten

Hier kann der Mittelwert $\bar{\varepsilon}$ aus der Formel

$$2\bar{\varepsilon} \approx P(\text{„ungleich“}|M)$$

bestimmt werden. Diese Wahrscheinlichkeit wird beim iterativen Abgleich geschätzt durch den Quotienten

$$\frac{\text{err}_X}{N_M},$$

wobei N_M die Anzahl der bisherigen erfolgreichen Matchvorgänge (also Treffer zwischen A und B) und err_X die Anzahl der dabei aufgetretenen Abweichungen in der Komponente X ist. Diese beiden Anzahlen sind also zu speichern und laufend anzupassen.

Solange im Verlauf der Match-Historie $\mathrm{err}_X = 0$ ist, wird ein vorzugebender Wert $\varepsilon_0 > 0$ verwendet (etwa $\varepsilon_0 = 0.05$).

Diese Überlegung lässt sich leicht auf den Fall verallgemeinern, dass die Fehlerraten konstant sind und in einem bekannten Verhältnis stehen. Sie passt zur Bestimmung globaler, aber auch werteabhängiger Gewichte, wenn man nur annimmt, dass die Fehlerraten konstant und gleich sind.

Newcombe [7, Section 9.1, Item 1] nimmt an, dass im allgemeinen Fall die wertespezifischen Gewichte für die *Ungleichheit* durch die globalen ersetzt werden können. Dann würde es reichen, stets nur die Summe $\bar{\varepsilon}_\alpha + \bar{\varepsilon}_\beta$ zu kennen, und die obige Schätzung ergibt ja gerade diese Summe. Das ist gerechtfertigt, wenn

- das betrachtete Merkmal wenigstens ungefähr gleichverteilt ist, siehe 2.4,
- oder die Anzahl der möglichen Ausprägungen $n = 2$ ist, siehe 2.6.

In anderen Fällen gibt es aber größere Abweichungen, die nicht tolerierbar sind; insbesondere kann man kaum annehmen, dass bei seltenen Werten die Fehlerrate geringer ist – wer „Pommerening“ statt „Müller“ heißt, erlebt das öfter –, und genau das bräuchte man, um die Quotienten

$$\frac{\varepsilon(x)}{p(x)}$$

einigermaßen konstant zu halten.

Werteabhängige Fehlerraten

Analog kann man auch werteabhängige Fehlerraten schätzen, wenn sie für beide Dateien gleich sind. Hierzu muss man nur die beiden Größen aus dem vorigen Abschnitt werteabhängig definieren:

$$\begin{aligned} N_M(x) &= \#\{a \in A \cap B \mid \xi(a) = x\} \approx p(x) \cdot \#(A \cap B) \approx p(x) \cdot N_M, \\ \mathrm{err}_X(x) &= \#\{a \in A \cap B \mid \xi(a) = x, \alpha(a) \neq x\} \approx p(x)\varepsilon(x) \cdot N_M. \end{aligned}$$

Also wird $\varepsilon(x)$ geschätzt durch den Quotienten

$$\varepsilon(x) \approx \frac{\mathrm{err}_X(x)}{N_M(x)} \approx \frac{\mathrm{err}_X(x)}{N_X(x)} \cdot \frac{N_B}{N_M}.$$

Die Anzahl $\mathrm{err}_X(x)$ ist also für jeden Wert $x \in X$ zu speichern und laufend anzupassen. Da die wahren Werte $\xi(a)$ nicht bekannt sind, zählt man ersatzweise so:

- Liegt ein Match vor mit $\alpha(a) = x$, $\beta(a) = x'$ und $x \neq x'$, so könnte x oder x' der richtige Wert $\xi(a)$ sein – vielleicht aber auch keiner von beiden. Da letzteres sehr unwahrscheinlich ist und die Entscheidung zwischen x und x' nicht getroffen werden kann, werden – da nach Annahme beide Werte mit gleicher Wahrscheinlichkeit falsch sind – $\mathrm{err}_X(x)$ und $\mathrm{err}_X(x')$ jeweils um $\frac{1}{2}$ erhöht.

Auch hier wird $\varepsilon(x)$ in der Anfangsphase, solange noch $\mathrm{err}_X(x) = 0$ ist, durch den vorgewählten Startwert ε_0 ersetzt. [Ist $\mathrm{err}_X(x) \neq 0$, so erst recht $N_X(x) \neq 0$, $N_B \neq 0$, $N_M \neq 0$.]

3.3 Schätzung der Fehlerraten im Modell II

Ein sehr leistungsfähiges Modell erhält man mit der Annahme, dass die Fehlerraten jeweils *zwei verschiedene Werte* annehmen können – es wird zwischen „sicheren“ und „unsicheren“ Daten mit einer geringen oder höheren Fehlerrate unterschieden. Dazu werden die abzugleichenden Dateien disjunkt zerlegt:

$$A = A_s \,\dot\cup\, A_u, \quad B = B_s \,\dot\cup\, B_u.$$

Die Fehlerraten auf diesen Teilen seien jeweils konstant:

- δ_s auf A_s,
- δ_u auf A_u,
- ε_s auf B_s,
- ε_u auf B_u.

Im Falle eines Matchs sind vier verschiedene Situationen zu unterscheiden und ergeben vier Gleichungen für die vier – als unbekannt angenommenen – Fehlerraten:

$$\begin{aligned}
\delta_s + \varepsilon_s &\approx P(\text{„ungleich“}|A_s \cap B_s),\\
\delta_u + \varepsilon_s &\approx P(\text{„ungleich“}|A_u \cap B_s),\\
\delta_s + \varepsilon_u &\approx P(\text{„ungleich“}|A_s \cap B_u),\\
\delta_u + \varepsilon_u &\approx P(\text{„ungleich“}|A_u \cap B_u).
\end{aligned}$$

Aus diesen können die linken Seiten geschätzt werden, wenn man die Zahl der jeweiligen Matches und die dabei vorkommenden Abweichungen wie in Abschnitt 3.2 nach den vier Situationen getrennt mitzählt; als Startwerte werden plausible Werte vorgegeben.

Leider sind die vier Gleichungen abhängig und gestatten nicht, die vier Fehlerraten einzeln zu bestimmen.

Hier hilft eine weitere realistische Modellannahme weiter:

Wir stellen uns vor, die mit s indizierten Teilmengen bestehen aus „sicheren“ Datensätzen, die aus einer (weitgehend) fehlerfreien Quelle stammen; zu denken ist hier etwa an Patienten-Stammdaten von einer Krankenkassenkarte. Die Fehlerrate ε_s beschreibt dann nur noch die Wahrscheinlichkeit, mit der sich der in der Datenbank B erfasste Wert geändert hat. Die Fehlerrate δ_s wird in der Annahme, dass der neu eingegebene Fall in A fehlerfrei ist und den aktuellen Wert enthält, $= 0$ gesetzt. Die Fehlerrate δ_u beschreibt den Fehler bei der Neueingabe, und ε_u bestimmt sich aus den in der Datenbank enthaltenen Eingabefehlern sowie der Änderungswahrscheinlichkeit.

Im folgenden wird also $\delta_s = 0$ gesetzt. Die übrigen Fehlerraten werden dann aus

$$\begin{aligned}
\varepsilon_s &\approx P(\text{„ungleich“}|A_s, B_s),\\
\delta_u &\approx P(\text{„ungleich“}|A_u, B_s) - \varepsilon_s,\\
\varepsilon_u &\approx P(\text{„ungleich“}|A_s, B_u)
\end{aligned}$$

bestimmt. Gezählt werden müssen:

- die Anzahl der Matches „sicher“/„sicher“, N^{ss}, und die Anzahl der dabei aufgetretenen Abweichungen, err_X^{ss} im Merkmal X,
- die Anzahl der Matches „unsicher“/„sicher“, N^{us}, und die Anzahl der dabei aufgetretenen Abweichungen, err_X^{us},
- die Anzahl der Matches „sicher“/„unsicher“, N^{su}, und die Anzahl der dabei aufgetretenen Abweichungen, err_X^{su}.

Die drei zu schätzenden Fehlerraten werden dann aus den Formeln

$$\varepsilon_s \approx \frac{\mathrm{err}_X^{ss}}{N^{ss}}, \quad \varepsilon_u \approx \frac{\mathrm{err}_X^{su}}{N^{su}}, \quad \delta_u \approx \frac{\mathrm{err}_X^{us}}{N^{us}} - \varepsilon_s \approx \frac{\mathrm{err}_X^{us}}{N^{us}} - \frac{\mathrm{err}_X^{ss}}{N^{ss}}$$

bestimmt, wobei für die Anfangsphase – solange im Verlauf der Match-Historie noch $\mathrm{err}_X^{xx} = 0$ ist – plausible vorgegebene Anfangswerte verwendet werden.

Die Gewichte werden dann nach den Formeln

$$\begin{aligned}
w(x,x) &= -\log p(x),\\
w(x,x') &= \begin{cases}
\log\left[\frac{\varepsilon_s}{p(x')}\right] - \log(n-1) & \text{für } x \in A_s, x' \in B_s, x \neq x',\\
\log\left[\frac{\varepsilon_u}{p(x')}\right] - \log(n-1) & \text{für } x \in A_s, x' \in B_u, x \neq x',\\
\log\left[\frac{\delta_u}{p(x)} + \frac{\varepsilon_s}{p(x')}\right] - \log(n-1) & \text{für } x \in A_u, x' \in B_s, x \neq x',\\
\log\left[\frac{\delta_u}{p(x)} + \frac{\varepsilon_u}{p(x')}\right] - \log(n-1) & \text{für } x \in A_u, x' \in B_u, x \neq x',
\end{cases}
\end{aligned}$$

berechnet.

Das geht natürlich nur, wenn nicht allzu lange nach Start des Verfahrens tatsächlich sichere Fälle auftreten. Was tun, wenn das nicht der Fall ist? In diesem Fall – solange nur unsichere Fälle vorgekommen sind – ist es wohl am besten, die Fehlerraten δ_u und ε_u als gleich anzunehmen und die Zählung und Schätzung aus Abschnitt 3.2 zu verwenden. Auf jeden Fall sollten also auch die werteabhängigen Abweichungen $\mathrm{err}_X(x)$ stets mitgezählt werden.

3.4 Benötigte Zähler

Nach den bisherigen Überlegungen sind beim iterativen Matchen also die folgenden Zähler mitzuführen:

Globale Zähler

- N_B = die Anzahl der bisher in der Datenbank B erfassten Datensätze. Sie wird immer dann erhöht, wenn ein neuer eingegebener Datensatz mit keinem der bisher erfassten Datensätze matcht (und auch nicht wegen offensichtlicher Mängel zurückgewiesen wird). N_B kann in den Gewichtsformeln immer als ≥ 1 angenommen werden, da bei leerer Datenbank kein Vergleich durchgeführt wird.

- N_M = die Anzahl der bisherigen positiv verlaufenen Match-Vorgänge. Darunter die Anzahlen

  - N^{ss}, wo beide Fälle „sicher“ waren,
  - N^{us}, wo ein „unsicherer“ Eingabefall zu einem „sicheren“ Datenbankfall passte,
  - N^{su}, wo ein „sicherer“ Eingabefall zu einem „unsicheren“ Datenbankfall passte.

  Solange N_M noch 0 ist, sind auch alle Zähler err_X und $\mathrm{err}_X(x)$ (siehe unten) 0, so dass dieser Fall hier nicht besonders behandelt zu werden braucht, da N_M nur für die Schätzung der Fehlerraten verwendet wird. Gleiches gilt jeweils für N^{xx} und err_X^{xx}.

Merkmalsabhängige Zähler

- n_X = die Anzahl der bisher in der Datenbank erfassten Ausprägungen des Merkmals X. Sie wird immer dann erhöht, wenn ein neuer aufgenommener Datensatz im Merkmal X eine bisher nicht vorhandene Ausprägung aufweist.

- err_X = die Anzahl der Abweichungen im Merkmal X bei den bisherigen positiv verlaufenen Match-Vorgängen. Darunter die Anzahlen

– err_X^{ss}, wo beide Fälle „sicher“ waren,

– err_X^{us}, wo ein „unsicherer“ Eingabefall zu einem „sicheren“ Datenbankfall passte,

– err_X^{su}, wo ein „sicherer“ Eingabefall zu einem „unsicheren“ Datenbankfall passte.

Die Behandlung des Anfangszustands $\mathrm{err}_X = 0$ bzw. $\mathrm{err}_X^{xx} = 0$ wird im nächsten Abschnitt 3.5 bei denjenigen Merkmalstypen beschrieben, wo sie eine Rolle spielt.

Ferner werden noch die merkmalsabhängigen Default-Fehlerraten ε_0, hier besser als ε_X bezeichnet, benötigt, die bei der Konfiguration des Match-Verfahrens als Konstanten gesetzt werden.

Werteabhängige Zähler

- $N_X(x)$ = die absolute Häufigkeit, mit der der Wert x bei den bisher erfassten Fällen im Merkmal X vorkommt.
- $\mathrm{err}_X(x) = \frac{1}{2}$-mal die Zahl der bisherigen positiven Matchvorgänge, bei denen genau einer der Fälle im Merkmal X den Wert x hatte.

Die Behandlung des Anfangszustands $N_X(x) = 0$ bzw. $\mathrm{err}_X(x) = 0$ wird im nächsten Abschnitt 3.5 bei den Merkmalstypen beschrieben, wo sie eine Rolle spielt.

3.5 Merkmalstypen

Vier Typen von Merkmalen werden unterschiedlich behandelt; diese sind hier nach zunehmender Komplexität aufgeführt:

(A) Das Merkmal ist gleichverteilt, insbesondere ist die Anzahl $n = n_X$ der Ausprägungen von vornherein bekannt und konstant. Auch die Fehlerrate wird als unabhängig vom Wert angenommen. Das Gewicht wird als globales Gewicht nach den Näherungsformeln für die Gleichverteilung aus 3.1 bestimmt, wobei im Fall $n \geq 6$ die vereinfachte Formel für „ungleich“ verwendet wird.

Das Gewicht $\log n_X$ der Gleichheit ist für diesen Merkmalstyp eine Konstante, die vorherberechnet werden kann. Gleiches gilt für den bei $n \leq 5$ nötigen Summanden

$$\Delta_X = \log\left(1 + \frac{1}{n-1}\right) = 0.30, 0.18, 0.12, 0.10 \quad \text{für } n = 2, 3, 4, 5$$

im Gewicht der Ungleichheit; für $n \geq 6$ wird $\Delta_X = 0$ gesetzt.

Der Anteil $\log \varepsilon$ des Gewichts der Ungleichheit hängt von der Summe der beiden Fehlerraten ab. Diese werden zu Beginn durch einen Schätzwert festgelegt und im Verlauf des Match-Vorgangs aufgrund der Ergebnisse laufend angepasst; der Algorithmus dafür wurde in 3.2 unter „globale Fehlerraten“ spezifiziert.

Die benötigten Zähler sind N_M und err_X, das hier als Schätzung für die Summe der beiden Fehlerraten verwendet wird, und die zu implementierende Formel ist:

$$\begin{aligned} w(\text{„gleich“}) &\approx \log n_X \quad \text{konstant}, \\ w(\text{„ungleich“}) &\approx \log \frac{\mathrm{err}_X}{N_M} + \Delta_X. \end{aligned}$$

Ist $\mathrm{err}_X = 0$, so ist $\frac{\mathrm{err}_X}{N_M}$ ist durch das vorgegebene ε_0 zu ersetzen. Ist $N_M = 0$, so erst recht $\mathrm{err}_X = 0$.

Beispiele: Geschlecht, Geburtsmonat.

(K) Das Merkmal beschreibt ein eindeutiges Schlüsselfeld. D. h., jeder Wert beschreibt ein eindeutiges Individuum, aber es ist nicht a priori bekannt, welche Werte auftreten, insbesondere auch nicht deren Anzahl. Darüber hinaus sind Änderungen oder Fehleingaben eines Werts möglich. Änderungen sind selten; bei einer Änderung sollte der alte Wert zusätzlich gespeichert bleiben.

Es ist ausreichend genau, ein solches Merkmal als gleichverteilt anzunehmen und wie Typ (A) zu behandeln, mit der einen Ausnahme, dass die Anzahl der Ausprägungen $= n_B$, der Anzahl bisher in der Datenbank erfassten Fälle, gesetzt wird. (Der Korrekturterm Δ_X aus (A) kann gleich auf 0 gesetzt werden.)

Beispiel: Krankenversichertennummer.

(B) Das Merkmal ist nicht gleichverteilt, aber die Wahrscheinlichkeit $p(x)$ der einzelnen Werte lässt sich a priori festlegen (bekannt oder genau genug schätzbar und stets > 0); insbesondere ist auch die Anzahl n_X bekannt. Die allgemeine Näherungsformel aus 3.1 für das werteabhängige Gewicht wird verwendet. Die Fehlerraten werden wie bei (A) behandelt, insbesondere als werteunabhängig angenommen. Die benötigten Zähler sind ebenfalls N_M und err_X, die zu implementierenden Formeln:

$$\begin{aligned} w(x,x) &\approx -\log p(x), \\ w(x,x') &\approx \log \left[\frac{1}{p(x)} + \frac{1}{p(x')} \right] + \log \frac{\mathrm{err}_X}{N_M} - \Delta_X \quad \text{für } x \neq x' \end{aligned}$$

mit dem konstanten Summanden $\Delta_X = \log(2n_X - 2)$.

Beispiel: Geburtstag.

(C) Die Wahrscheinlichkeit der einzelnen Werte lässt sich nicht a priori festlegen; sie muss dann aus der relativen Häufigkeit in der Datenbank der bisher erfassten Fälle geschätzt werden. Als Formel ist die allgemeine Näherungsformel aus 3.1 für das werteabhängige Gewicht zu verwenden. Es werden alle Zähler aus 3.4 benötigt.

Damit werden die Formeln im Modell I zu

$$\begin{aligned} w(x,x) &\approx \log N_B - \log N_X(x), \\ w(x,x') &\approx \log\left[\frac{\mathrm{err}_X(x)}{N_X(x)^2} + \frac{\mathrm{err}_X(x')}{N_X(x')^2}\right] + 2\log N_B - \log N_M \\ &\quad - \log(n_X - 1) \quad \text{für } x \neq x'. \end{aligned}$$

Hierbei ist x' der Wert aus der bestehenden Datenbank und x der Wert des neuen Falles. Ist $N_M = 0$, so auch $\mathrm{err}_X(x') = 0$ und $\mathrm{err}_X(x) = 0$; ist $N_X(x) = 0$, so auch $\mathrm{err}_X(x) = 0$, analog für x'. Die gesondert zu behandelnden Fälle in der Anfangsphase sind also das Verschwinden der werteabhängigen Fehlerzähler. Hier wird im Fall von x die Fehlerrate $\varepsilon_\alpha(x) = \varepsilon_0$ und die Wahrscheinlichkeit $p(x) = 1/2N_B$ gesetzt, also der Quotient

$$\frac{\varepsilon_\alpha(x)}{p(x)} = 2N_B\varepsilon_0.$$

Im Fall von x' ist jedenfalls $N_X(x') \geq 1$, und man kann

$$\frac{\varepsilon_\alpha(x')}{p(x')} = \frac{N_B\varepsilon_0}{N_X(x')}$$

setzen.

Außerdem muss man am Anfang des Match-Vorgangs eine Zeitlang auch mit dem Fall $n_X = 1$ rechnen; in diesem Fall wird n_X durch $\frac{3}{2}$ ersetzt.

Im Modell II gilt eine analoge Überlegung mit Behandlung der Fehlerraten nach 3.3.

Beispiele: Geburtsjahr, Wohnort.

(D) Das Merkmal ist ein Komplex aus mehreren abhängigen Merkmalen [...]

Beispiel: Name.

3.6 Verhalten der Zähler beim Matchen

Der Abgleichvorgang beim iterativen Matchen besteht jeweils daraus, dass ein neuer Fall mit dem Bestand der Datenbank verglichen wird.

Unentschiedener Ausgang

Der neue Fall kann bereits vor dem Match-Versuch wegen offensichtlicher Mängel abgewiesen werden; in diesem Fall ist in der Datenbank nichts weiter zu ändern. Ferner kann beim Matchversuch das Ergebnis „unentschieden“ herauskommen; auch in diesem Fall ist nichts zu ändern.

Negativer Ausgang

Ergibt der Matchversuch keinen möglichen Treffer, so ist der neue Fall in die Datenbank aufzunehmen. Folgende Zähler sind zu ändern:

- N_B wird inkrementiert.
- Für jedes Merkmal vom Typ (C) wird, sofern der Wert x in der Eingabe nicht fehlt („missing value“),
  - der Wertezähler $N_X(x)$ inkrementiert,
  - falls $N_X(x) = 0$ war, d. h., der Wert x für das Merkmal X bisher nicht vorkam, der Ausprägungszähler n_X inkrementiert.

Die Zähler N_M, N^{ss}, N^{us}, N^{su}, $\mathrm{err}_X \ldots$ bleiben ungeändert.

Positiver Ausgang

Wesentlich komplizierter ist die Behandlung der Zähler, wenn der Matchversuch erfolgreich ist, d. h., genau einen Treffer liefert. Hier bleibt N_B ungeändert, aber N_M wird inkrementiert. Für das weitere sind vier Fälle zu unterscheiden:

(ss) Sowohl der neue Fall als auch der passende „alte“ Fall in der Datenbank sind als „sicher“ markiert. Dann wird N_{ss} inkrementiert. Für jedes Merkmal X wird geprüft:

- Falls der neue Wert x mit dem alten Wert x' übereinstimmt oder fehlt, passiert nichts weiter.
- Falls der alte Wert fehlte, aber der neue x vorhanden ist, wird dieser in die Datenbank aufgenommen. Falls das Merkmal vom Typ (C) ist, wird $N_X(x)$ inkrementiert; falls außerdem der Wert x bisher nicht vorkam, wird n_X inkrementiert.
- Falls beide Werte vorhanden, aber unterschiedlich sind, wird dies als Änderung des wahren Wertes von x' nach x interpretiert. Es wird err_X^{ss} inkrementiert. Falls das Merkmal vom Typ (C) ist, wird außerdem
  - $N_X(x')$ dekrementiert; wird es dabei zu 0, wird zusätzlich n_X dekrementiert;

- $N_X(x)$ inkrementiert; war es vorher 0, wird zusätzlich n_X inkrementiert;
- $\mathrm{err}_X(x)$ und $\mathrm{err}_X(x')$ um jeweils $\frac{1}{2}$ erhöht. [*Achtung*: Es kann also danach $N_X(x') = 0$, aber $\mathrm{err}_X(x') > 0$ sein.]

(us) Der neue Fall ist als „unsicher“, der passende Fall aus der Datenbank als „sicher“ markiert. Zunächst wird N_{us} inkrementiert. Dann wird für jedes Merkmal X geprüft:

- Falls der neue Wert x mit dem alten Wert x' übereinstimmt oder fehlt, passiert nichts weiter.
- Falls der alte Wert fehlte, aber der neue x vorhanden ist, wird dieser in die Datenbank aufgenommen. [*Schlecht!* Wird der Datenbankfall dadurch unsicher?] Falls das Merkmal vom Typ (C) ist, wird $N_X(x)$ inkrementiert; falls außerdem der Wert x bisher nicht vorkam, wird n_X inkrementiert.
- Falls beide Werte vorhanden, aber unterschiedlich sind, wird der Wert x' aus der Datenbank als der korrekte angenommen und daher nicht verändert. Es wird err_X^{us} inkrementiert. Falls das Merkmal vom Typ (C) ist, werden außerdem
  - $\mathrm{err}_X(x)$ und $\mathrm{err}_X(x')$ um jeweils $\frac{1}{2}$ erhöht. [*Achtung*: Es kann also danach $N_X(x') = 0$, aber $\mathrm{err}_X(x') > 0$ sein.]

(su) Der neue Fall ist als „sicher“, der passende Fall aus der Datenbank als „unsicher“ markiert. Zunächst wird N_{su} inkrementiert. Dann wird für jedes Merkmal X geprüft:

- Falls der neue Wert x mit dem alten Wert x' übereinstimmt oder fehlt, passiert nichts weiter.
- Falls der alte Wert fehlte, aber der neue x vorhanden ist, wird dieser in die Datenbank aufgenommen. Falls das Merkmal vom Typ (C) ist, wird $N_X(x)$ inkrementiert; falls außerdem der Wert x bisher nicht vorkam, wird n_X inkrementiert.
- Falls beide Werte vorhanden, aber unterschiedlich sind, wird der neu eingegebene Wert x als der korrekte angenommen und daher der alte Wert x' durch x ersetzt. Es wird err_X^{su} inkrementiert. Falls das Merkmal vom Typ (C) ist, wird außerdem
  - $N_X(x')$ dekrementiert; wird es dabei zu 0, wird zusätzlich n_X dekrementiert;
  - $N_X(x)$ inkrementiert; war es vorher 0, wird zusätzlich n_X inkrementiert;
  - $\mathrm{err}_X(x)$ und $\mathrm{err}_X(x')$ um jeweils $\frac{1}{2}$ erhöht. [*Achtung*: Es kann also danach $N_X(x') = 0$, aber $\mathrm{err}_X(x') > 0$ sein.]

(uu) Der neue Fall und der passende Fall aus der Datenbank sind als „unsicher“ markiert. [N_{uu} wird nicht mitgeführt.] Für jedes Merkmal X wird geprüft:

- Falls der neue Wert x mit dem alten Wert x' übereinstimmt oder fehlt, passiert nichts weiter.
- Falls der alte Wert fehlte, aber der neue x vorhanden ist, wird dieser in die Datenbank aufgenommen. Falls das Merkmal vom Typ (C) ist, wird $N_X(x)$ inkrementiert; falls außerdem der Wert x bisher nicht vorkam, wird n_X inkrementiert.
- Falls beide Werte vorhanden, aber unterschiedlich sind, wird der alte Wert x' als der bessere angenommen und daher nicht verändert. [err_X^{su} gibt's nicht.] Falls das Merkmal vom Typ (C) ist, werden außerdem
  - $\mathrm{err}_X(x)$ und $\mathrm{err}_X(x')$ um jeweils $\frac{1}{2}$ erhöht. [*Achtung*: Es kann also danach $N_X(x') = 0$, aber $\mathrm{err}_X(x') > 0$ sein.]

3.7 Behandlung typischer Merkmale

Geschlecht

Dieses Merkmal ist vom Typ (A) mit $n_X = 2$, also $\Delta_X = 0.30$. Als Gewichte verwendet man daher:

$$\begin{aligned} w(\text{„gleich“}) &= 0.30, \\ w(\text{„ungleich“}) &= 0.30 + \begin{cases} \log \varepsilon_0, & \text{wenn } \mathrm{err}_X = 0, \\ \log \frac{\mathrm{err}_X}{N_M}, & \text{wenn } \mathrm{err}_X > 0. \end{cases} \end{aligned}$$

Geburtsmonat

Auch dieses Merkmal kann als Typ (A) angenommen werden mit $n_X = 12$, also $\Delta_X = 0$. Als Gewichte werden verwendet

$$\begin{aligned} w(\text{„gleich“}) &= 1.08, \\ w(\text{„ungleich“}) &= \begin{cases} \log \varepsilon_0, & \text{wenn } \mathrm{err}_X = 0, \\ \log \frac{\mathrm{err}_X}{N_M}, & \text{wenn } \mathrm{err}_X > 0. \end{cases} \end{aligned}$$

Geburtstag

Hier ist, wie in 2.5 gesehen, der Typ (B) anzunehmen mit $n_X = 31$ und $\Delta_X = \log 60 = 1.78$. Die Wahrscheinlichkeiten $p(x)$ für $x = 1, \ldots, 31$ sind wie in 2.5 anzusetzen.

Geburtsjahr

Dieses Merkmal ist *nicht* gleichverteilt; die Verteilung hängt sehr stark von der betrachteten Population ab. Hier muss man also Typ (C) annehmen.

Vorname

Beim Vornamen ist die Unabhängigkeitsannahme fraglich; die Verteilung der Vornamen ist vom Geburtsjahr abhängig, aber auch vom Familiennamen („Jürgen Bauer“, „Jesco von Puttkamer“). Solche eher schwachen Abhängigkeiten werden in der Praxis aber vernachlässigt. Natürlich besteht aber eine sehr starke Abhängigkeit vom Geschlecht; bei Verwendung des Vornamens kann man das Geschlecht praktisch unberücksichtigt lassen.

Je nachdem, ob zum Vornamen auch phonetische Codes gebildet werden, nimmt man als Typ (C) oder (D) an.

Name

Dies ist ein sehr kompliziertes Merkmal. Im GPOH-PID-Dienst wird nach dem Muster der Krebsregistrierung in Rheinland-Pfalz [12] ein komplexes Modell für Beobachtungsfunktion genommen: Sei X die Menge der Zeichenketten $\leq N$, etwa ≤ 100. Die datenerzeugenden Prozesse sind dann Funktionen

$$\alpha\colon A \longrightarrow X, \quad \beta\colon B \longrightarrow X.$$

Die Vergleichsfunktion wird aus zwei Schritten zusammengesetzt:

1. Schritt. Normalisierung: Der Name wird normalisiert (Umlaute auflösen, Großbuchstaben) und in bis zu drei Bestandteile zerlegt – weitere Bestandteile werden verworfen; Namenszusätze kommen, sofern dort nichts anderes steht, in den dritten Bestandteil.

$$\text{Name} \mapsto (N_1, N_2, N_3).$$

2. Schritt. Es werden zwei verschiedene phonetische Codes gebildet: Φ_K (Kölner Phonetik nach Postel [10] [11]) und Φ_H (Hannoveraner Phonetik nach Michael [6]).

Damit ist die Vergleichsfunktion

$$\begin{aligned} A \times B &\xrightarrow{\gamma} \Gamma = X^{10}, \\ (\text{Name}, \text{Name'}) &\mapsto \begin{pmatrix} N_1 & N_2 & N_3 & \Phi_K(\text{Name}) & \Phi_H(\text{Name}) \\ N_1' & N_2' & N_3' & \Phi_K(\text{Name'}) & \Phi_H(\text{Name'}) \end{pmatrix}. \end{aligned}$$

Je nachdem, ob Gleichheit oder Ähnlichkeit in einem oder beiden phonetischen Codes besteht, wird man unterschiedliche Gewichte erhalten; die genauere Bestimmung steht noch aus.

3.8 Festlegung der Schwellenwerte

[. . . fehlt noch . . .]

Literatur

[1] AUTOMATCH *User's Manual.* 1997.

[2] Thomas R. Belin, Donald B. Rubin: A method for calibrating false-match rates in record linkage. J. Amer. Stat. Ass. 90 (1995), 694–707.

[3] Ivan P. Fellegi, Alan B. Sunter: A theory for record linkage. J. Amer. Stat. Ass. 64 (1969), 1183–1210.

[4] M. Fortini, B. Liseo, A. Nuccitelli, M. Scanu: On Bayesian record linkage. ??? ?? (???), ???–???.

[5] Matthew A. Jaro: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa. J. Amer. Stat. Ass. 84 (1989), 414–420.

[6] J. Michael: Doppelgänger gesucht – Ein Programm für kontextsensitive phonetische Textumwandlung. c't 25/1999, 252–261.

[7] Howard B. Newcombe: *Handbook of Record Linkage.* Oxford Univ. Press 1988.

[8] Howard B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James: Automatic linkage of vital records. Science 130 (1959), 954–959.

[9] Howard B. Newcombe, Martha E. Fair, Pierre Lalonde: Concepts and practices that improve probabilistic linkage. Statistics Canada Symposium, November 1987, 127–138.

[10] H.-J. Postel: Die Kölner Phonetik – Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. IBM-Nachrichten 19 (1969), 925–931.

[11] H.-J. Postel: Probleme beim Aufbau eines Informationssystems für Sicherheitsbehörden (II). Datenverarbeitung in Steuer, Wirtschaft und Recht, 2/1975, 55–61.

[12] I. Schmidtmann, H.-J. Appelrath, J. Michaelis, W. Thoben: Empfehlungen an die Bundesländer zur technischen Umsetzung der Verfahrensweisen gemäß Gesetz über Krebsregister (KRG). Informatik, Biometrie und Epidemiologie in Medizin und Biologie 27 (1996), 101–110.

[13] William E. Winkler: The State of Record Linkage and Current Research Problems. Technical Report, U. S. Bureau of the Census, ca 1999. [Online im WWW: `http://census.gov/...`]