

# Testumfang für die Ermittlung und Angabe von Fehlerraten in biometrischen Systemen

Peter Unruh  
SRC Security Research & Consulting GmbH  
peter.unruh@src-gmbh.de

## Einleitung

Biometrische Systeme werden durch zwei wichtigen Parameter charakterisiert: FAR (False Acceptance Rate) und FRR (False Rejection Rate). Dabei hängen diese Fehlerraten von einem Schwellwert ab, der in der meisten Systemen einstellbar ist und als Kriterium für die JA/NEIN-Entscheidung verwendet wird. Der Schwellwert sagt aus, wie viel Prozent der Merkmale übereinstimmen sollen, um vom biometrischen System „erkannt“ zu werden. Maßgeblich für die Wahl des Schwellwertes und dadurch auch der Fehlerraten sind die Anforderungen der geplanten Anwendung. Steht die Sicherheit im Vordergrund, wird der Schwellwert so hoch gewählt, dass möglichst wenig, im Idealfall keine Falschakzeptanzen vorkommen. Die FAR ist in diesem Fall klein. Wird bei der Anwendung mehr Wert auf Komfort gelegt als auf die Sicherheit, so sorgt der nicht so hoch gesetzte Schwellwert dafür, dass möglichst wenig Zurückweisungen von Berechtigten stattfinden. Die FRR ist in diesem Fall klein.

## Aufgabenstellung

Die Frage, die sich jeder Hersteller von biometrischen Systemen stellt, lautet: „Wie viele Testversuche reichen aus, um mit Hilfe einer statistischen Methode zu beweisen, dass die angegebenen Fehlerraten bei einem festgelegten Schwellwert mit großer Wahrscheinlichkeit einen bestimmten Toleranzwert nicht überschreiten?“

Dabei kann für die FAR der Toleranzwert  $\delta$  wichtig sein, um neben dem typischen Wert  $FAR_{\text{TYP}}$  die Angabe für  $FAR_{\text{MAX}} = FAR + \delta$  mit großer Wahrscheinlichkeit garantieren zu können.  $FAR - \delta$  ist für diese Problematik unkritisch, weil das System eine bessere als typisch angegebene Fehlerrate aufweist.

## Statistische Signifikanz

In der Sprache der Statistik lautet die oben gestellte Frage folgendermaßen:

Wie groß muss der Stichprobenumfang  $n$  gewählt werden, damit bei festem Signifikanzniveau  $\alpha$  eine vorgegebene Abweichung  $\delta$  zwischen dem arithmetischen Mittel  $\bar{x}$  der Stichprobe ( $FAR_{\text{MAX}}$ ) und dem angenommenen Sollwert  $\mu_0$  ( $FAR_{\text{TYP}}$ ) gerade noch als signifikant erkannt wird (damit wäre  $FAR_{\text{MAX}}$  „verbindlich“). Abweichungen kleiner als  $\delta$  können praktisch als unbedeutend angesehen werden und brauchen nicht nachgewiesen werden.

Mit anderen Worten soll die Wahrscheinlichkeit, dass die Abweichung  $\bar{x} - \mu_0$  größer als  $\delta$  ist, den Wert  $\alpha$  (Irrtumswahrscheinlichkeit) nicht überschreiten:  $P(\bar{x} - \mu_0 \geq \delta) \leq \alpha$ .

Bei der Herleitung des Stichprobenumfangs wird von Eigenschaften und Funktionen der Normalverteilung ausgegangen, deshalb werden sie hier aufgeführt [STAT]:

Die stetige Zufallsgröße  $X$ , die alle Werte der reellen Zahlen zwischen  $-\infty$  und  $+\infty$  annehmen kann, genügt der Normalverteilung, wenn ihre Dichte durch

$$\varphi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{für } -\infty < x < +\infty \text{ gegeben ist.}$$

Der Erwartungswert  $\mu$  und die Streuung (oder Varianz)  $\sigma^2$  heißen Parameter der Normalverteilung  $N(\mu; \sigma^2)$ . Die zu der oben beschriebenen Dichte gehörende Verteilungsfunktion ist

$$\Phi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

Werden für die Parameter  $\mu = 0$  und  $\sigma^2 = 1$  gesetzt, so ist die standardisierte Form der Normalverteilung  $N(0;1)$  gegeben. Dichte und Verteilungsfunktion sind dann:

$$\varphi(x; 0, 1) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

$$\Phi(x; 0, 1) = \Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Für beide Funktionen gibt es Tabellen, aus denen für vorgegebene Werte von  $x$  die Dichte  $\varphi(x)$  und die Verteilungsfunktion  $\Phi(x)$  abgelesen werden kann.

Ist  $X$  eine normalverteilte Zufallsgröße mit den Parametern  $\mu$  und  $\sigma^2$ , so ist die Zufallsgröße  $Z = \frac{X - \mu}{\sigma}$  normalverteilt mit den Parametern 0 und 1. Für jede reelle Zahl  $x$  gilt dann

$$\varphi(x; \mu, \sigma^2) = \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}\right),$$

$$\Phi(x; \mu, \sigma^2) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Dies ermöglicht die Verwendung der Tabellen für die standardisierten Normalverteilung  $N(0;1)$  zur Berechnung der Dichte und der Verteilungsfunktion einer nach  $N(\mu; \sigma^2)$  normalisierten Zufallsgröße  $X$  mit beliebigen Parametern  $\mu$  und  $\sigma^2$ .

Unter der Annahme, dass  $n$  Stichproben  $x_i$  genommen werden, so dass  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , kann für  $\bar{x}$  von einer Normalverteilung  $N(\mu; \frac{\sigma^2}{n})$  ausgegangen werden. Dabei genügt die Stichprobenfunktion  $Z = \frac{\bar{x} - \mu}{\sigma} \sqrt{n}$  der standardisierten Normalverteilung  $N(0;1)$  [WAHR].

Die Wahrscheinlichkeit  $P$  der Beziehung  $P(\bar{x} - \mu_0 \geq \delta) \leq \alpha$  kann folgendermaßen umgeschrieben werden:  $P(\bar{x} - \mu_0 \geq \delta) = P(\frac{\bar{x} - \mu_0}{\sigma} \sqrt{n} \geq \frac{\delta}{\sigma} \sqrt{n}) = P(Z \geq \frac{\delta}{\sigma} \sqrt{n})$ .

Die Beziehung  $P(Z \geq \frac{\delta}{\sigma} \sqrt{n}) \leq \alpha$  kann durch  $P(Z < \frac{\delta}{\sigma} \sqrt{n}) \geq 1 - \alpha$  ersetzt werden.

Aus [WAHR] ist zu entnehmen, dass  $P(Z < \frac{\delta}{\sigma} \sqrt{n}) = \Phi(\frac{\delta}{\sigma} \sqrt{n})$  und  $\Phi(Z_{1-\alpha}) = 1 - \alpha$ , deshalb kann die Beziehung  $P(\bar{x} - \mu_0 \geq \delta) \leq \alpha$  als  $\Phi(\frac{\delta}{\sigma} \sqrt{n}) \geq 1 - \alpha$ , oder als  $\Phi(\frac{\delta}{\sigma} \sqrt{n}) \geq \Phi(Z_{1-\alpha})$  aufgeschrieben werden.

Für die Ermittlung des Stichprobenumfangs  $n$  wird nun die Ungleichung  $\frac{\delta}{\sigma} \sqrt{n} \geq Z_{1-\alpha}$  verwendet:

$$n \geq \frac{Z_{1-\alpha}^2 \sigma^2}{\delta^2}$$

Mit Hilfe dieser Formel kann der Testumfang bestimmt werden, wenn bei den Tests als Ergebnis ein Wert mit bestimmter Varianz  $\sigma^2$  (oder bekannter Standardabweichung  $\sigma$ ) zu erwarten ist. Der Wert  $Z_{1-\alpha}$  wird aus der Tabelle der Verteilungsfunktion  $\Phi(Z_{1-\alpha})$  ausgelesen.

Bei der Ermittlung der FAR werden Tests durchgeführt, deren Ergebnisse bei festgelegten Schwellwert sich entweder als „akzeptiert“ oder als „abgewiesen“ zusammenfassen lassen. Dadurch ist eine Binomialverteilung vorgegeben. Wenn die Wahrscheinlichkeit eines Ergebnisses mit  $p$  bezeichnet wird, dann kann der Erwartungswert  $\mu$  und die Streuung  $\sigma^2$  entsprechend durch  $np$  und  $np(1-p)$  definiert werden [STAT].

Die Stichprobenfunktion ist für diesen Fall:

$$Z = \frac{x - np}{\sqrt{np(1-p)}} = \frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Für unsere Fragestellung ist es wichtig, dass die Wahrscheinlichkeit, dass die Abweichung zwischen relativer Häufigkeit  $\frac{x}{n}$  und der Wahrscheinlichkeit  $p$  größer ist als  $\delta$ , unter der Irrtumswahrscheinlichkeit  $\alpha$  liegt:  $P(\frac{x}{n} - p \geq \delta) \leq \alpha$

Die Wahrscheinlichkeit  $P$  der Beziehung  $P(\frac{x}{n} - p \geq \delta) \leq \alpha$  kann folgendermaßen umgeschrieben werden:

$$P(\frac{x}{n} - p \geq \delta) = P(\frac{\frac{x}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{\delta}{\sqrt{\frac{p(1-p)}{n}}}) = P(Z \geq \frac{\delta}{\sqrt{\frac{p(1-p)}{n}}}).$$

Die Beziehung  $P(Z \geq \frac{\delta}{\sqrt{\frac{p(1-p)}{n}}}) \leq \alpha$  kann durch  $P(Z < \frac{\delta}{\sqrt{\frac{p(1-p)}{n}}}) \geq 1 - \alpha$  ersetzt werden.

Da  $P(Z < \frac{\delta}{\sqrt{\frac{p(1-p)}{n}}}) = \Phi(\frac{\delta}{\sqrt{\frac{p(1-p)}{n}}})$  und  $\Phi(Z_{1-\alpha}) = 1 - \alpha$ ,

kann die Beziehung  $P(\frac{x}{n} - p \geq \delta) \leq \alpha$  als  $\Phi(\frac{\delta}{\sqrt{\frac{p(1-p)}{n}}}) \geq 1 - \alpha$  oder als

$\Phi(\frac{\delta}{\sqrt{\frac{p(1-p)}{n}}}) \geq \Phi(Z_{1-\alpha})$  aufgeschrieben werden.

Für die Ermittlung des Stichprobenumfangs  $n$  kann nun die Ungleichung  $\frac{\delta}{\sqrt{\frac{p(1-p)}{n}}} \geq Z_{1-\alpha}$

verwendet werden:

$$n \geq \frac{Z_{1-\alpha}^2 p(1-p)}{\delta^2}.$$

Für die Irrtumswahrscheinlichkeit  $\alpha$  werden in der Praxis folgende Werte verwendet:

0,05, 0,01 und 0,001. Mit Hilfe der Tabelle der Verteilungsfunktion  $\Phi(Z_{1-\alpha})$  wird der entsprechende Wert  $Z_{1-\alpha}$  bestimmt:

$\alpha$	0,05	0,01	0,001
$Z_{1-\alpha}$	1,645	2,326	3,090

In der folgenden Tabelle ist für jede Irrtumswahrscheinlichkeit  $\alpha$  abhängig von der FAR und der Abweichung  $\delta$  der Testumfang  $n$  angegeben:

FAR <sub>TYP</sub>	$\delta$	FAR <sub>MAX</sub>	$n_{\alpha=0,05}$	$n_{\alpha=0,01}$	$n_{\alpha=0,001}$
0,05	0,01	0,06	1.285	2.570	4.535
0,05	0,005	0,055	5.141	10.280	18.141
0,05	0,001	0,051	128.536	256.988	453.535
0,04	0,01	0,05	1.039	2.078	3.666
0,04	0,005	0,045	4.156	8.310	14.666
0,04	0,001	0,041	103.911	207.755	366.647
0,03	0,01	0,04	787	1.574	2.778
0,03	0,005	0,035	3.150	6.298	11.114
0,03	0,001	0,031	78.745	157.439	277.850
0,02	0,01	0,03	530	1.060	1.871
0,02	0,005	0,025	2.122	4.242	7.486
0,02	0,001	0,021	53.038	106.041	187.143
0,01	0,01	0,02	268	536	945
0,01	0,005	0,015	1.072	2.142	3.781
0,01	0,001	0,011	26.790	53.562	94.526
0,005	0,001	0,006	13.462	26.916	47.502
0,005	0,0005	0,0055	53.850	107.664	190.007
0,005	0,0001	0,0051	1.346.247	2.691.612	4.750.180
0,001	0,001	0,002	2.703	5.405	9.539
0,001	0,0005	0,0015	10.813	21.619	38.154
0,001	0,0001	0,0011	270.332	540.487	953.855
0,0001	0,0001	0,0002	27.058	54.097	95.471
0,0001	0,00005	0,00015	108.230	216.389	381.886
0,0001	0,00001	0,00011	2.705.754	5.409.735	9.547.145
0,00001	0,00001	0,00002	270.600	541.022	954.800
0,00001	0,000005	0,000015	1.082.399	2.164.089	3.819.202
0,00001	0,000001	0,000011	27.059.979	54.102.219	95.480.045

Vergleichbare Ergebnisse bekommt man, wenn die Regeln "Rule of 3" und "Rule of 30" angewandt werden. Diese Regeln werden z. B. in [BEM] und [BPTEST] zur Ermittlung des Testumfangs vorgeschlagen, wenn der Vergleichsalgorithmus eines biometrischen Systems getestet werden soll. Der Unterschied im Vergleich zum Testen auf Systemebene besteht darin, dass die realen Bedingungen sowie die Systemumgebung nicht berücksichtigt werden. Für die Tests werden biometrische Daten verwendet, die entweder früher vom System erfasst und gespeichert wurden oder möglicherweise durch Verwendung anderer Systeme erzeugt wurden. Analog zu FAR ist beim Testen des Vergleichsalgorithmus der Parameter FMR (False Match Rate) aus Sicherheitssicht von entscheidender Bedeutung.

Mit "Rule of 3" kann die folgende Frage beantwortet werden: "Welche minimale Fehlerrate kann bei  $N$  unabhängigen Vergleichen statistisch begründet werden?" Die Wahrscheinlichkeit, dass die nach der "Rule of 3" ermittelte Fehlerrate  $p \approx \frac{3}{N}$  fehlerhaft ist und bei  $N$  Vergleichsversuchen eine fehlerhafte Übereinstimmung vorkommt, ist kleiner als 5%.

Mit anderen Worten, wenn bei  $N$  Vergleichsversuchen keine fehlerhafte Übereinstimmung zu Stande kam, dann kann mit großer Wahrscheinlichkeit (95%) garantiert werden, dass die FMR kleiner als  $\frac{3}{N}$  ist.

Die folgende Tabelle enthält einige auf diese Weise ermittelte Zahlenbeispiele:

$N$	60	300	3.000	30.000
$p$	0,05	0,01	0,001	0,0001

Die "Rule of 30" stammt von Doddington [RULE30] und bietet folgendermaßen Hilfe für die Ermittlung der Testumfänge:

Um mit 90%-er Sicherheit behaupten zu können, dass die „echte“ Fehlerrate von der bei Vergleichen festgestellten Fehlerrate höchstens um  $\pm 30\%$  abweicht, müssen mindestens 30 fehlerhafte Übereinstimmungen beobachtet werden.

Die Relation ist nicht linear, mindestens 260 fehlerhafte Übereinstimmungen sind zu beobachten, um eine Höchstabweichung von  $\pm 10\%$  zu garantieren.

Wenn es beispielsweise zu 30 fehlerhaften Übereinstimmungen innerhalb von 3.000 Versuchen gekommen ist, dann liegt die „echte“ FMR zwischen 0,007 und 0,013. Um für einen Algorithmus eine FMR von  $0,001 \pm 10\%$  angeben zu können, müssen 260 fehlerhafte Übereinstimmungen beobachtet werden. Dies bedeutet, dass 260.000 Vergleiche durchgeführt wurden.

Dabei wird vorausgesetzt, dass die Vergleichsversuche statistisch unabhängig sind. Streng genommen sind für das letzte Beispiel 260.000 verschiedene biometrische Personendaten erforderlich, die mit einer Referenz verglichen werden.

Die Forderung der statistischen Unabhängigkeit hat zur Folge, dass die Anzahl der Testpersonen sich verdoppeln muss, wenn sowohl die FAR als auch die FRR statistisch korrekt zu ermitteln sind.

In der Praxis wird oft auf die statistische Unabhängigkeit verzichtet, um die Anzahl der Testpersonen zu verringern. Dabei kann die Irrtumswahrscheinlichkeit nicht so präzise (wie oben) angegeben werden. Angenommen, es nehmen  $N$  Personen teil, deren biometrische Daten erfasst und als Referenzdaten gespeichert werden. Bei gegenseitigem Vergleich der erfassten Daten mit Referenzdaten sind  $N(N-1)/2$  Vergleichsversuche möglich.

Mit 250 Testpersonen können etwa 30.000 Vergleichsversuche durchgeführt werden, um die FMR zu ermitteln. Wird keine Übereinstimmung festgestellt, so kann die FMR im besten Fall nach "Rule of 3" den Wert 0,0001 erhalten.

## Referenzen

- [STAT] Statistische Methoden und ihre Anwendungen, Erwin Kreyszig, 4., unveränd. Nachdr. der 7. Auflage, 1991
- [WAHR] Wahrscheinlichkeitsrechnung mathematische Statistik und statistische Qualitätskontrolle, Regina Storm, 7. Auflage, 1979
- [BEM] Common Criteria, Common Methodology for Information Technology Security Evaluation, Biometric Evaluation Methodology Supplement, Common Criteria Biometric Evaluation Methodology Working Group, Version 1.0, August 2002
- [BPTEST] Best Practices in Testing and Reporting Performance of Biometric Devices, By A. J. Mansfield, National Physical Laboratory and J. L. Wayman, San Jose State University, Version 2.0, August 2002
- [RULE30] The NIST speaker recognition evaluation: Overview methodology, systems, results, perspective. Doddington, G. R., Przybocki, M. A., Martin, A. F., and Reynolds, D. A., Speech Communication, 2000, 31(2-3), 225-254