

Maschinelles Lernen zur Hautkrebsvorhersage



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Eine Bachelorarbeit von Daniel Fischer

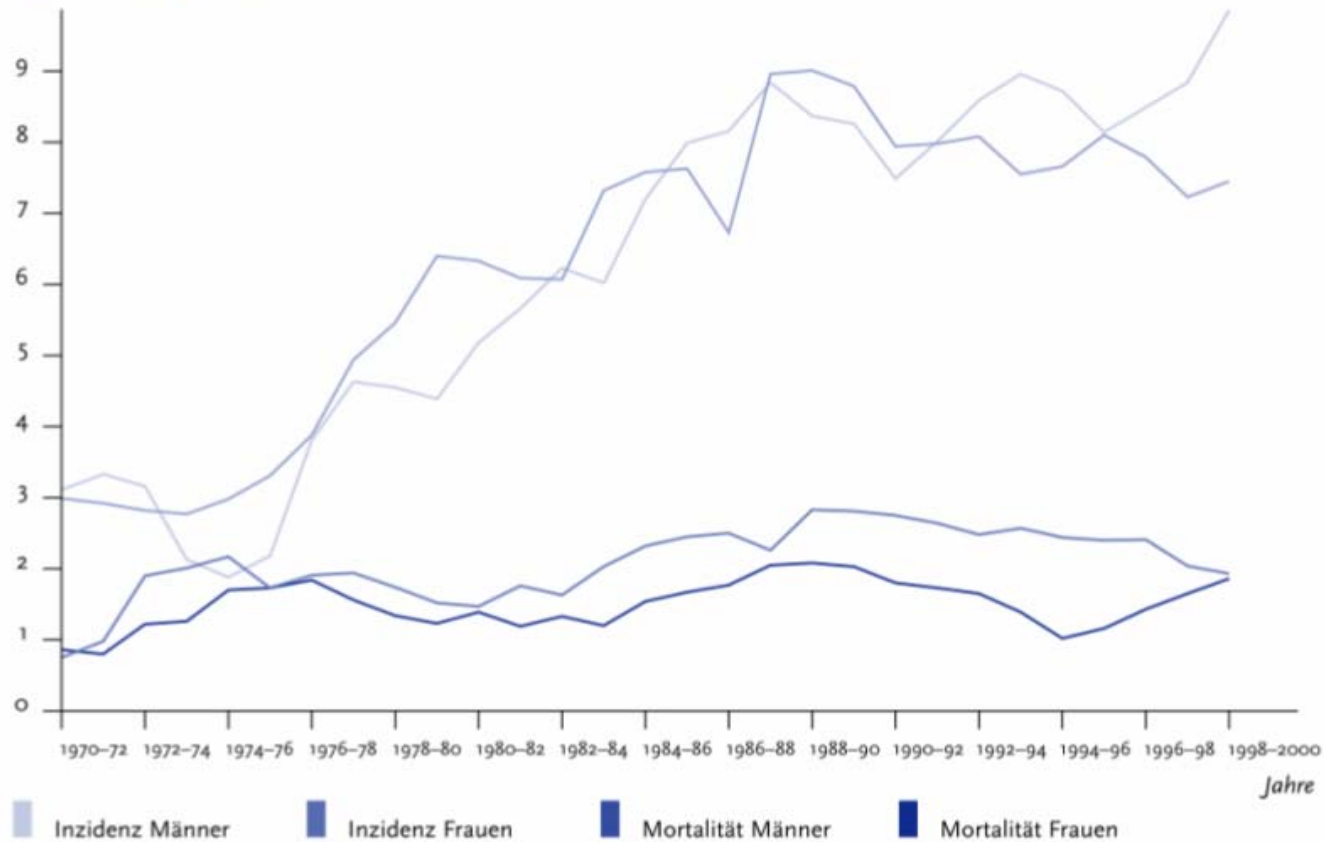
Betreuung: Dipl. Inf. Frederik Janssen

Prof. Dr. Johannes Fürnkranz

Dr. med. Matthias Herbst

Motivation

Alterstandardisierte Erkrankungsraten und Mortalität am malignen Melanom der Haut,
Saarland 1970–2000, gleitende Mittelwerte
Angaben pro 100.000 Einwohner
Quelle: Krebsregister Saarland



[1]

Ziele der Arbeit

- **Ausgangslage:** Wissen über Risikofaktoren (s. Anhang)

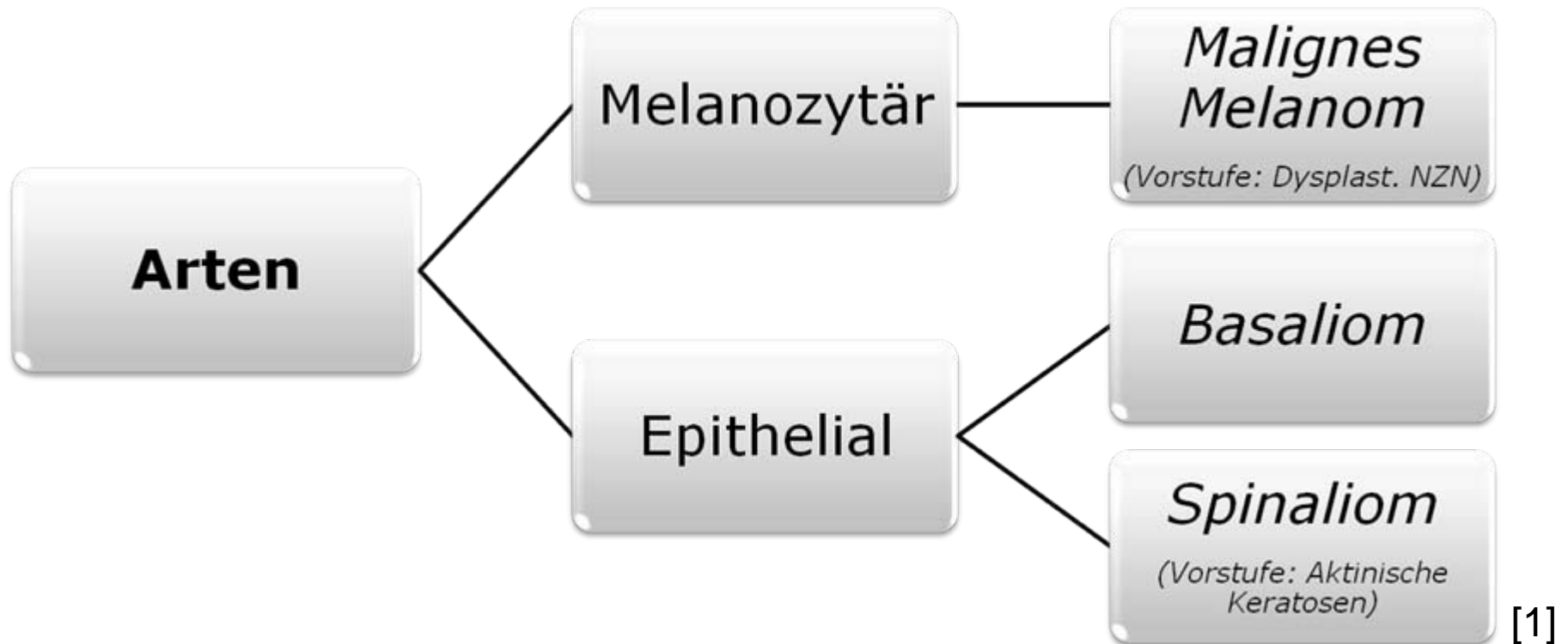
(1) Verifikation von Erkenntnissen

(2) Erstellung eines performanten Klassifikationssystems für Patienten

Gliederung

1. Einführung in das Thema Hautkrebs
2. Herkunft der Daten
3. Grundlagen des Data Mining
4. Data Preprocessing
 1. Kodierung
 2. Konvertierung
 3. Bereinigung von Inkonsistenzen
 4. Behandlung fehlender Werte
 5. Feature Subset Selection
5. Algorithmen
6. Experimente
7. Diskussion und Ausblick

1. Einführung in das Thema Hautkrebs



[1]

2. Herkunft der Daten

- „Hautcheck-Programm“ der Qualitätsgemeinschaft südhessischer Dermatologen e.V.
- Erhebung von 6.938 Patientendaten (Fragebogen)
 - Persönliche Daten
 - Ärztliche Diagnosen
- Finanzierung: BKK Merck KG
- Dienstleister: Iatrocon GmbH

2. Herkunft der Daten

▪ Patienten-Fragebogen:

Vom Teilnehmer auszufüllen:

1. Alter in Jahren _____ 2. Geschlecht: 2.1 weiblich 2.2 männlich

3. Fragen zum Freizeitverhalten. Wie oft halten Sie sich bei intensiver Sonneneinstrahlung in der Sonne auf?
 3.1 So häufig wie möglich 3.2 Gelegentlich 3.3 eher selten 3.4 Ich melde die Sonne

4. Wie reagiert Ihre Haut auf 30 Minuten Besonnung ohne Vorbereitung?
 4.1 Immer Sonnenbrand / niemals Bräunung 4.2 Häufig Sonnenbrand / schwache Bräunung
 4.3 Selten Sonnenbrand / gute Bräunung 4.4 Sehr selten Sonnenbrand / sehr gute Bräunung
 4.5 Keine sichtbaren Reaktionen, da braune Haut 4.6 Keine sichtbaren Reaktionen, da schwarze Haut

5. Wie viele schwere Sonnenbrände (schmerzhaft mit Blasen) haben Sie in Ihrem Leben erlitten?

| Gruppe | Sonnenbrand | | | |
|--------------------------------|-------------|--------|--------|-----|
| | nie | selten | häufig | oft |
| 5.1 Kind (0 bis 8 J.) | | | | |
| 5.2 Jugendlicher (8 bis 16 J.) | | | | |
| 5.3 Erwachsener (ab 16. J.) | | | | |

6. Schützen Sie sich vor Sonneneinstrahlung?
 6.1 konsequent angewendet 6.2 selten/sporadisch angewendet 6.3 keine Aussage

7. Benutzen Sie ein Solarium?
 7.1 2-3 x pro Woche 7.2 1 x pro Woche 7.3 Selten 7.4 Nie

8. Welche Sportarten betreiben Sie?
 8.1 Fußball 8.3 Hockey 8.5 Segeln 8.4 Schwimmen
 8.5 Tennis 8.6 Reiten 8.7 Radsport 8.8 Leichtathletik
 8.9 Rudern 8.10 Kanu 8.11 Joggen 8.12 Wandern
 8.13 _____

9. Hatten Sie schon mal Hautkrebs?
 9.1 Ja 9.2 Nein

10. Ist in Ihrer Familie Hautkrebs aufgetreten?
 10.1 Ja 10.2 Nein 10.3 Unbekannt

2. Herkunft der Daten

■ Ärzte-Fragebogen:

Vom Arzt auszufüllen:

1. Anzahl der Pigmentmale?

1.1 Bis zu 10 1.2 10 bis 20
 1.3 20 bis 50 1.4 Mehr als 50

2. Dysplast. NZN ? 2.1

3. Praecancerosen (Ak) ? 3.1

4. Basaliom (Bcc) ? 4.1 →

5. Spinaliom (Sc) ? 5.1 →

6. Melanom (MM) ? 6.1 →

7. Wird eine Therapie eingeleitet? 7.1 Nein 7.2 Ja →

7.2.1 Op erforderlich 7.2.2 PDT
 7.2.3 Lokaltherapie 7.2.4 Hautpflege

Histo.: _____

8. Beurteilung!

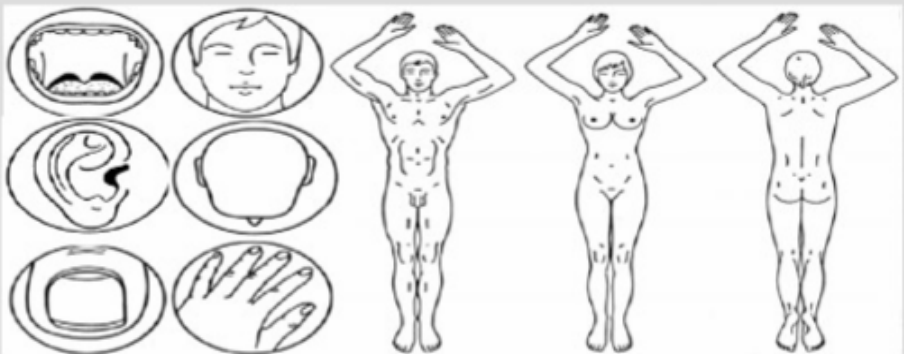
8.1 Erhöhtes Risiko
 8.2 Kein erhöhtes Risiko

9. Wiedervorstellung!

9.1 Alle 2-3 Jahre 9.2 Pro Jahr 1 mal 9.3. Pro Jahr 2 mal

10. Aufklärung wurde durchgeführt!

10.1 Hauttyp 10.2 Lichtschutz 10.3 ABCD - Regel MM



Praxisstempel

Bezeichnung
Krankenkasse:

Datum: _____
Unterschrift Arzt: _____

Interne
Patienten-Nr.: _____

Arzt-Nr.: _____
Bogen-Nr.: _____
wird von Clearingstelle ausgefüllt

3. Grundlagen des Data Mining

- **Motivation:**

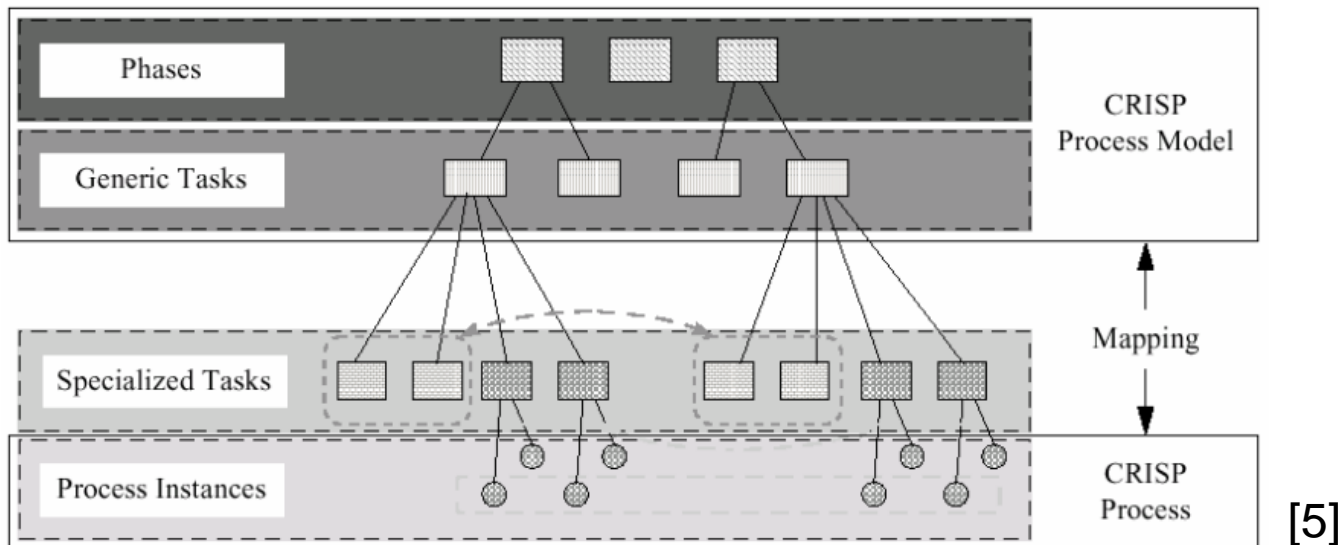
„We are drowning in information but starving for knowledge.“
(Rutherford D. Roger)

- **Definition:**

„Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.“ [3]

3. Grundlagen des Data Mining

- Leitfaden zur Erstellung der Arbeit
- Viele Alternativen (Fayyad, Han & Kamber etc.)
- CRISP (1996-1999): Quasi-Industriestandard [4]



4. Data Preprocessing

- **Motivation:** Datenbasis häufig *verzerrt, unvollständig, inkonsistent*
- **Ziel:** Qualitäts- und Effizienzsteigerung
- **Aufgaben:**
 - Kodierung / Skalierung der Daten (Kapitel 4.1)
 - Konvertierung der Daten: CSV → ARFF (Kapitel 4.2)
 - Bereinigung von Inkonsistenzen (Kapitel 4.3)
 - Behandlung fehlender Werte (Kapitel 4.4)
 - Feature Subset Selection (Kapitel 4.5)

4.1 Kodierung der Daten

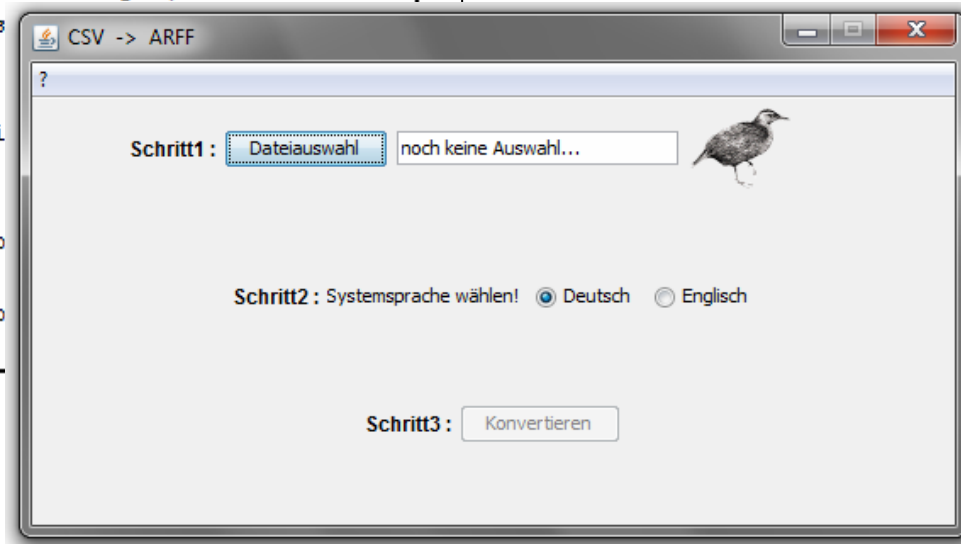
- Unterschiedliche Skalenniveaus → Mapping erforderlich
- 3 Attributarten auf Fragebogen → Index (s, AUSPRÄGUNGSNAME, -)

| | A | B | C | D | E | F | G | H | I |
|----|------------|-------|----------|-------|------------|------------|--------------|--------------|------------|
| 1 | Datum | Arzt# | Bogen# | Alter | Geschlecht | Geschlecht | Outdoor-Zeit | Outdoor-Zeit | Malignes M |
| 2 | - | - | - | - | männlich | weiblich | häufig | gelegentlich | s |
| 3 | 26.10.2009 | 102 | 10010750 | 55 | | x | | x | |
| 4 | 26.10.2009 | 102 | 10010752 | 52 | x | | | x | |
| 5 | 26.10.2009 | 102 | 10010747 | 40 | | x | | x | |
| 6 | 26.10.2009 | 102 | 10010749 | 58 | x | | | x | |
| 7 | 26.10.2009 | 102 | 10010748 | 65 | x | | x | | x |
| 8 | 26.10.2009 | 102 | 10010746 | 62 | x | | | x | |
| 9 | 26.10.2009 | 102 | 10010751 | 60 | x | | | x | |
| 10 | 26.10.2009 | 10 | 10010753 | 52 | | x | | | |
| 11 | 26.10.2009 | 10 | 10010754 | 49 | x | | | x | |
| 12 | 26.10.2009 | 10 | 10010755 | 23 | x | | | x | |
| 13 | 24.10.2009 | 10 | 10010756 | 64 | x | | | | |
| 14 | 24.10.2009 | 10 | 10010757 | 54 | | x | | x | |
| 15 | 24.10.2009 | 10 | 10010758 | 37 | | x | | | |
| 16 | 24.10.2009 | 10 | 10010759 | 52 | | x | | | |

4.2 Konvertierung der Daten

```
1 @relation skin_cancer
2
3 @attribute name {Bob,Alice,Eve,Trudy,Oscar}
4 @attribute gender {male,female}
5 @attribute age numeric
6 @attribute 'outdoor sports' {yes,no}
7 @attribute cancer {yes
8
9
10 % Kommentare werden mi
11 @data
12 Bob,male,29,no,yes
13 Alice,female,33,yes,no
14 Eve,female,30,no,no
15 Trudy,female,40,yes,no
16 Oscar,male,15,yes,yes
```

} *Header*



4.3 Bereinigung von Inkonsistenzen

- Reihe von Untersuchungsergebnissen (Pigmente)
- Sport-Arten → zusammengeführt
- Inkonsistenzen in „Beurteilung“

| | | <i>Beurteilung</i> | | | |
|-------------------------|----------------|--------------------|----------------------|----------------|------|
| | | erhöhtes Risiko | kein erhöhtes Risiko | fehlender Wert | |
| <i>malignes Melanom</i> | ja | 26 | 3 | 2 | 31 |
| | nein | 2505 | 4189 | 213 | 6907 |
| | fehlender Wert | 0 | 0 | 0 | 0 |
| | | 2531 | 4192 | 215 | |

4.4 Behandlung fehlender Werte

- Gründe des Fehlens erörtern
- **Ersetzungsmethoden:**
 - Löschen aller unvollständigen Instanzen
 - Beschaffung der fehlenden Werte
 - Globalwert-Ersetzung
 - Klassifikation der fehlenden Werte:
 - Lineare Regression
 - Multiple Lineare Regression
 - K-Nearest Neighbor Verfahren
- **Mehrere Trainingsmengen:**
 - Fehlende Werte T_1
 - Löschung T_2
 - Volle Ersetzung T_3
 - Halbe Ersetzung T_4

4.4 Behandlung fehlender Werte

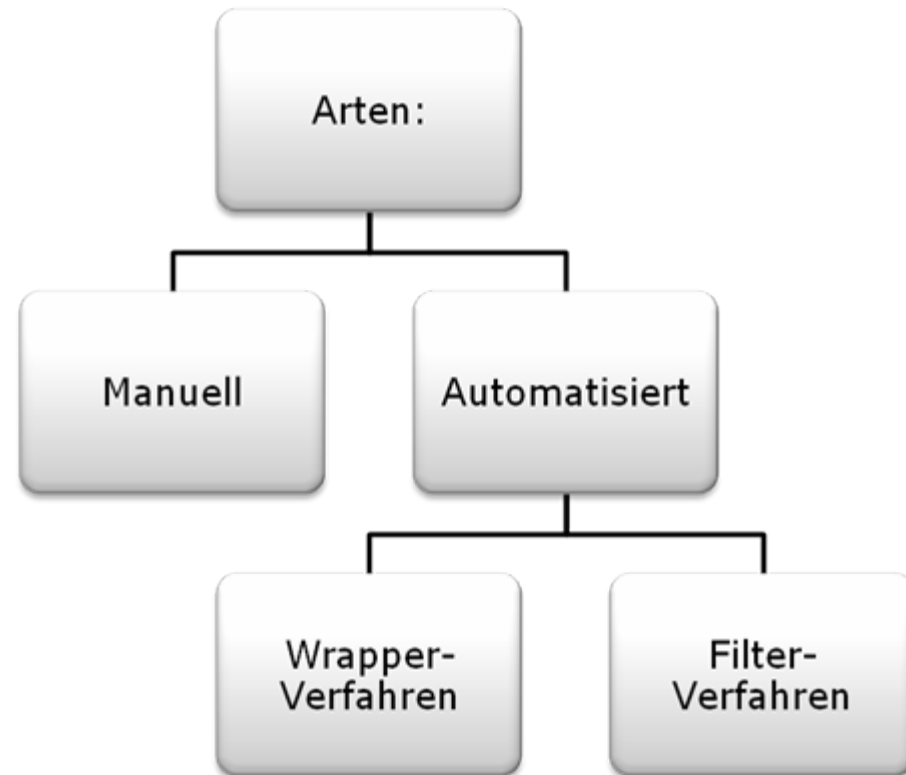
Ersetzte Attributwerte:

- Sonnenbrand als Kind (2.035)
- Sonnenbrand als Jugendlicher (1.754)
- Sonnenbrand als Erwachsener (1.064)
- Outdoorzeit bei intensiver Sonneneinstrahlung (93)
- Solariumnutzung (239)
- Hautreaktion (180)
- Hautkrebs in der Eigenanamnese (158)

Halbe Ersetzung T_4

4.5 Feature Subset Selection

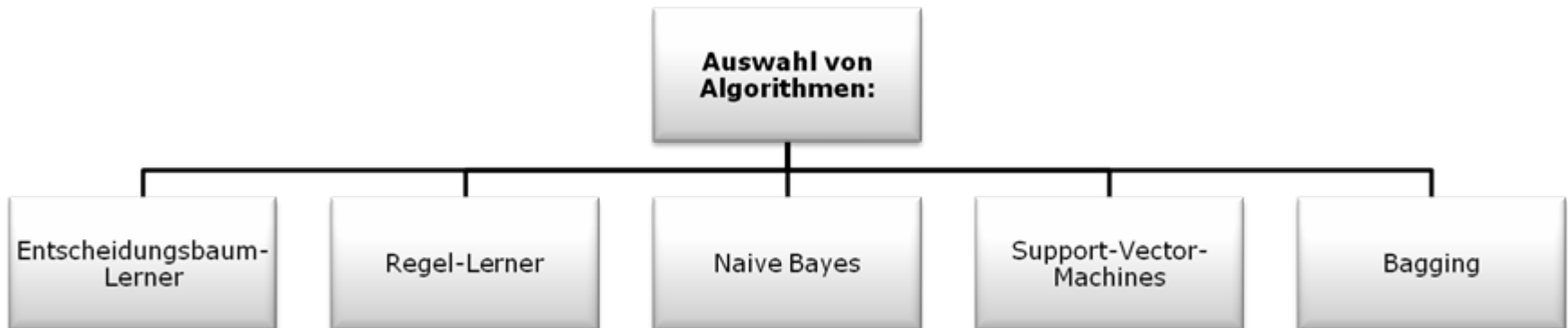
- **Ausgangslage:** Häufige Auswahl irrelevanter Attribute („Arzt#“)
- **Idee:** Auswahl einer relevanten Teilmenge von Attributen (Feature Subset)
- **Ziele:** Qualitätssteigerung der Klassifikation



[6]

5. Algorithmen

- Vielzahl von Klassifikationsalgorithmen mit individuellen Stärken und Schwächen



6. Experimente

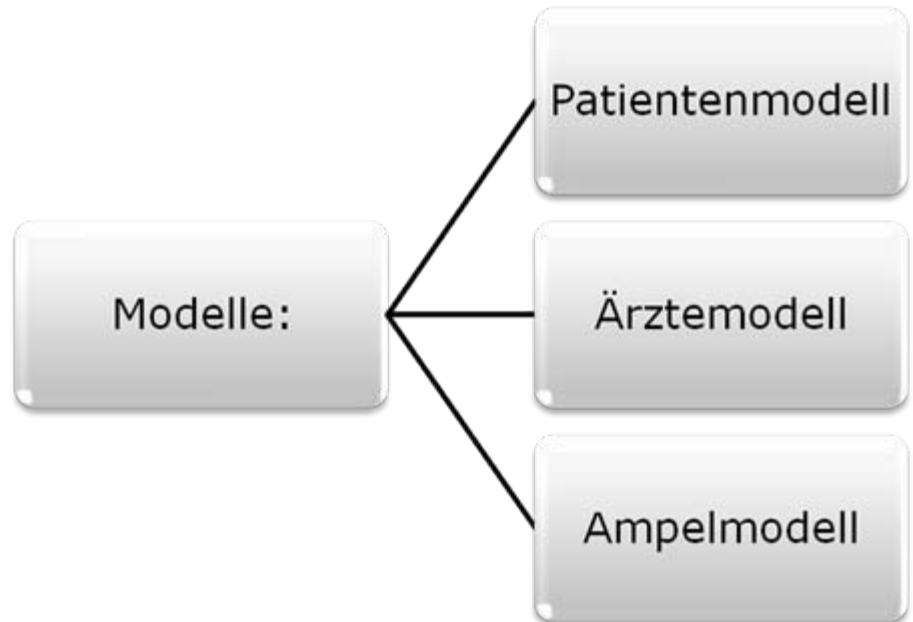
■ Unterschiedliche Anwendungsgebiete der Modelle



Malignes Melanom + Spinaliom

Basaliom + Aktinische Keratosen

Kein Befund



| | <i>Basaliom</i> | <i>Spinaliom</i> | <i>malignes Melanom</i> |
|--------------------------------|-----------------|------------------|-------------------------|
| <i>Metastasierungswahrsch.</i> | 0,2515 | 5,5 | ∞ |
| <i>Letalität</i> | 0,1 | 5 | 17,5 |

[1]

6. Experimente

| Patienten- modell | <i>größter Informationszuwachs</i> | | | | | |
|-----------------------------|------------------------------------|-------------|------------|----------------|--------------------|--------|
| | <i>Baseline (Erwartungswert)</i> | <i>JRip</i> | <i>J48</i> | <i>SMO</i> | <i>Naive Bayes</i> | |
| <i>T1 (mit fehlenden)</i> | 62.353% | 64.257% | 64.495% | 65.031% | 64.168% | 2.677% |
| <i>T2 (Löschung)</i> | 61.252% | 62.949% | 63.124% | 63.922% | 63.323% | 2.670% |
| <i>T3 (volle Ersetzung)</i> | 62.353% | 64.302% | 63.573% | 65.035% | 63.707% | 1.949% |
| <i>T4 (halbe Ersetzung)</i> | 62.605% | 63.988% | 63.686% | 65.025% | 63.794% | 2.420% |
| <i>Durchschnitt</i> | | 63.874% | 63.719% | 64.753% | 63.748% | |

6. Experimente

| Ärztmodell | <i>Baseline (Erwartungswert)</i> | <i>JRip</i> | <i>J48</i> | <i>SMO</i> | <i>Naive Bayes</i> | <i>größter Informationszuwachs</i> |
|-----------------------------|----------------------------------|----------------|----------------|------------|--------------------|------------------------------------|
| <i>T1 (mit fehlenden)</i> | 62.353% | 74.163% | 74.148% | 73.554% | 73.152% | 11.795% |
| <i>T2 (Löschung)</i> | 61.252% | 74.376% | 74.027% | 73.927% | 72.954% | 13.124% |
| <i>T3 (volle Ersetzung)</i> | 62.353% | 73.881% | 74.253% | 73.554% | 72.810% | 11.900% |
| <i>T4 (halbe Ersetzung)</i> | 62.605% | 74.876% | 74.249% | 74.379% | 73.493% | 12.271% |
| Durchschnitt | | 74.324% | 74.169% | 73.853% | 73.102% | |

6. Experimente

| | <i>Baseline (Erwartungswert)</i> | <i>JRip</i> | <i>J48</i> | <i>SMO</i> | <i>Naive Bayes</i> | <i>größter Informationszuwachs</i> |
|-----------------------------|----------------------------------|----------------|----------------|------------|--------------------|------------------------------------|
| <i>T1 (mit fehlenden)</i> | 87.712% | 87.668% | 87.963% | 87.172% | 87.579% | 0.251% |
| <i>T2 (Löschung)</i> | 88.872% | 89.047% | 89.122% | 88.872% | 88.997% | 0.249% |
| <i>T3 (volle Ersetzung)</i> | 87.669% | 87.874% | 88.130% | 87.669% | 87.446% | 0.461% |
| <i>T4 (halbe Ersetzung)</i> | 89.134% | 89.501% | 89.177% | 89.134% | 89.134% | 0.367% |
| <i>Durchschnitt</i> | | 88.522% | 88.598% | 88.212% | 88.289% | |

7. Diskussion und Ausblick



- Informationszuwachs bestätigt Anwendungsmöglichkeit
- Fragebogen-Konzeption → Verzerrung
 - *„Frage 3: Wie oft halten Sie sich bei intensiver Sonneneinstrahlung in der Sonne auf?“*
 - a) *So häufig wie möglich*
 - b) *Gelegentlich*
 - c) *Eher selten*
 - d) *Ich meide die Sonne*
- Inkonsistenzen in „Beurteilung“ → Verzerrung
- Geringe Anzahl an Befunden (Sparsity) → Sampling etc.
- False-Negatives → Kostensensitives Lernen

**Vielen Dank für die
Aufmerksamkeit!**

Literaturverzeichnis

- [1] Breitbart, E., Wende, A., Mohr, P., Greinert, R., & Volkmer, B. (2004). Hautkrebs (Robert-Koch-Institut, Hrsg.) *Gesundheitsberichterstattung des Bundes* (22).
- [2] Hossiep, R., & Wottawa, H. (1993). Die Angewandte Psychologie in Schlüsselbegriffen. In R. Hossiep, H. Wottawa, & A. Schorr (Hrsg.), *Handwörterbuch der Angewandten Psychologie* (S. 131-136). Bonn: Deutsche Psychologien Verlags GmbH.
- [3] Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, MA: MIT Press
- [4] KDnuggets. (2007). *Data Mining Methodology*. Abgerufen am 23. November 2010 von http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm
- [5] Chapman, P., Clinton, J., Kerber R., Khabaza, T., Reinartzs, T., Shearer, C., et al. (2000). *CRISP-DM 1.0*. Abgerufen am 23. November 2010 von <http://www.crisp-dm.org/CRISPWP-0800.pdf>
- [6] Liu, H., & Yu, L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. (I. C. Society, Hrsg.) *IEEE Transactions on Knowledge and Data Engineering*, 17(4), S. 491-500

Literaturverzeichnis

-
- [7] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1), S. 81-106
- [8] Cohen, W. W. (1995). Fast Effective Rule Induction. *In Proceedings of the Twelfth International Conference on Machine Learning*, S. 115-123.