

Päsentation Bachelorarbeit

Michael Bräunlein

mbraeunlein@gmail.com

Betreuer: Prof. J. Fürnkranz



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Klassifikation der Schwierigkeitsgrade von Sudokus mit Methoden des maschinellen Lernens



- ▶ Einleitung
- ▶ Lösungsmethoden
- ▶ Featurevektoren
- ▶ J48 Klassifizierer
- ▶ Ergebnisse



- ▶ Sudokus finden sich überall
- ▶ Unterschiedliche Bewertungsskalen
- ▶ Unterschiedliche Einteilungsverfahren
- ▶ Bisher kein Verfahren zur Einteilung mit maschinellem Lernen
- ▶ Sudokus sind zur Bearbeitung mit Computern prädestiniert



- ▶ Sudoku hat nur eine Regel
- ▶ In jeder Zeile, jeder Spalte und jedem Block muss jede Ziffer von 1 bis 9 genau einmal vorkommen
- ▶ Jedes Sudoku hat eine eindeutige Lösung
- ▶ Das Sudoku gilt dann als gelöst, wenn alle Felder ausgefüllt sind



- ▶ Lösungsmethoden sagen viel über den Schwierigkeitsgrad aus
- ▶ Jeder Spieler benutzt Lösungsmethoden
- ▶ Kandidatenlisten erleichtern das Finden von Zahlenkonstellationen, die Voraussetzung für bestimmte Lösungsmethoden sind
- ▶ Es gibt viele verschiedene Lösungsmethoden, grob werden zwei Kategorien unterschieden

| | | |
|--------|--------|-------------|
| 2 7 | 1 9 | 2 5 7 |
| 4 | 8 | 3 |
| 2 7 | 6 9 | 2 7 |

Hidden Single



TECHNISCHE
UNIVERSITÄT
DARMSTADT

| | | | | | | | | |
|----------|------------|----------|----------|------------|----------|------------|-----------------|---------------|
| 2 5 | 2 5 | 2 | 1 | 4 | 2 | 6 | 3 | 2 |
| 7 9 | 8 9 | 7 8 | | 7 | | | | 8 9 |
| 1 2 3 | 1 2 | 1 2 | 5 | 2 3 | 2 6 | 4 | 7 | 1 2 |
| 6 9 | 8 9 | 8 | | | | | | 8 9 |
| 1 2 3 | 1 2 | 1 2 | 2 3 | 2 3 | 9 | 1 2 | 2 | 5 |
| 7 | 6 4 | 4 | 7 8 | 7 | | | 8 | |
| 2 | 3 | 2 | 2 | 1 | 2 5 | 8 | 2 5 6 | 4 |
| 7 9 | | 7 | 7 9 | 7 | | | | |
| 1 2 | 1 2 | 6 | 4 | 8 | 3 | 9 | 2 5 | 2 |
| 7 | | | | | | | | 7 |
| 4 | 2 | 5 | 2 | 6 | 2 | 2 3 | 1 | 2 3 |
| | 8 9 | | 7 9 | 7 | | 7 | | 7 |
| 8 | 1 2 4 5 | 1 2 4 | 6 | 2 3 7 9 | 1 2 7 | 1 2 3 5 | 2 4 5 | 1 2 3 7 9 |
| 1 2 5 | 7 | 3 | 2 9 | 2 9 | 8 | 1 2 5 | 2 4 5 6 9 | 1 2 6 9 |
| 1 2 | 6 | 9 | 2 3 7 | 5 | 4 | 1 2 3 7 | 2 8 | 1 2 3 7 8 |

Pointing Pair / Triple



TECHNISCHE
UNIVERSITÄT
DARMSTADT

| | | | | | | | | |
|---------------|-----------|-----------|---------|---------|---------|---------|-----------|---------|
| 4 5 6 | 7 | 3 | 1 | 8 | 6 9 | 4 | 5 6 9 | 2 |
| 1 | 9 | 8 | 2 | 4 6 7 | 5 | 4 7 | 6 7 | 3 |
| 4 5 6 | 4 5 6 | 2 | 4 6 7 9 | 6 7 9 | 3 | 4 7 8 9 | 5 6 7 8 9 | 1 |
| 4 5 6 8 | 3 | 4 5 6 8 9 | 6 8 9 | 6 9 | 7 | 1 | 2 | 4 5 8 9 |
| 2 5 8 | 2 5 8 | 7 | 3 | 1 | 4 | 6 | 5 8 9 | 5 8 9 |
| 2 4 6 8 | 1 | 9 | 5 | 2 6 | 2 6 8 | 4 7 8 | 3 | 4 7 8 |
| 2 4 5 6 7 8 9 | 2 4 5 6 8 | 1 | 4 6 8 9 | 4 5 6 9 | 2 6 8 9 | 3 | 6 7 8 9 | 6 7 8 9 |
| 3 | 5 6 8 | 5 6 | 7 | 5 6 9 | 1 | 2 | 4 | 6 8 9 |
| 2 4 6 7 8 9 | 2 4 6 8 | 4 6 | 3 | 2 6 8 9 | 5 | 1 | 6 7 8 9 | 6 |

Two-String-Kite



TECHNISCHE
UNIVERSITÄT
DARMSTADT

| | | | | | | | |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 4 ³ | 2 | 6 | 7 ¹ | 8 ⁵ | 1 ⁴ | 9 ¹ | 3 ⁵ |
| 1 | 9 | 5 ⁸ | 4 | 2 ⁵ | 3 | 2 ⁷ | 2 ⁵ |
| 4 ³ | 5 ⁸ | 7 | 1 ⁵ | 9 ⁶ | 2 ⁶ | 1 ² | 2 ³ |
| 3 ¹ | 3 ⁶ | 9 | 2 ⁶ | 8 | 5 | 1 ² | 4 |
| 5 | 4 | 2 | 3 | 7 | 1 | 6 | 8 |
| 7 | 1 ⁶ | 1 ⁸ | 9 | 4 | 2 ⁶ | 3 | 1 ² |
| 9 | 7 | 3 | 1 ² | 6 | 4 | 5 | 8 |
| 2 ⁶ | 1 ⁵ | 1 ⁵ | 8 | 3 | 9 | 2 ⁴ | 7 |
| 2 ⁶ | 6 ⁸ | 4 | 1 ² | 7 | 9 | 3 | 1 ² |

XY-Wing



TECHNISCHE
UNIVERSITÄT
DARMSTADT

| | | | | | | | | |
|----------------------|----------------------|----------------------|--------------------------|------------------------|--------------------------|----------------------|---|--------------------|
| 5 | 1 | 9 | 4 | 7 | 6 | 8 | 2 | 3 |
| <small>7 8</small> | 6 | 2 | <small>3 5 9</small> | <small>5 8 9</small> | <small>3 8</small> | <small>7 9</small> | 1 | 4 |
| <small>7 8</small> | 4 | 3 | <small>2 9</small> | 1 | <small>2 8 9</small> | 5 | 6 | <small>7 9</small> |
| 2 | <small>3 9 7</small> | <small>6 7</small> | <small>3 6 9</small> | <small>6 9</small> | 5 | 4 | 8 | 1 |
| 1 | 5 | <small>4 6 7</small> | <small>2 3 6 4 7</small> | <small>2 6 8</small> | <small>2 3 4 7 8</small> | <small>3 6 7</small> | 9 | <small>6 7</small> |
| <small>4 6</small> | <small>3 9</small> | 8 | 1 | <small>4 6 7 9</small> | <small>3 7 9</small> | <small>3 6 7</small> | 5 | 2 |
| <small>6 9</small> | 7 | 5 | <small>2 6 9</small> | <small>2 4 6 9</small> | <small>2 4 9</small> | 1 | 3 | 8 |
| 3 | 8 | 1 | <small>5 6 7 9</small> | <small>5 6 7 9</small> | <small>7 9</small> | 2 | 4 | <small>6 9</small> |
| <small>4 6 9</small> | 2 | <small>4 6</small> | 8 | 3 | 1 | <small>6 9</small> | 7 | 5 |

Was sind Featurevektoren



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ Merkmalsvektor
- ▶ n-dimensionaler Vektor
- ▶ Repräsentation eines Objekts
- ▶ Ein Eintrag steht für eine Eigenschaft des beschriebenen Sudokus
- ▶ Merkmalsvektoren sind die Eingabe des Klassifikationsalgorithmus

Wie werden Featurevectoren erzeugt



TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ Am Anfang bekannte Zahlen
- ▶ Einträge der Kandidatenlisten
- ▶ Hinzugefügte Zahlen
- ▶ Entfernte Zahlen
- ▶ Unterschiedliche Lösungswege für Sudokus möglich
- ▶ Einfachster Lösungsweg gesucht
- ▶ Ungelöste Felder
- ▶ Insgesamt 261 Features



- ▶ Fast gleiche Sudokus mit vertauschten Zahlen
- ▶ Gleicher Schwierigkeitsgrad
- ▶ Unterschiedliche Featurevektoren bei gleichem Lösungsweg



- ▶ Fast gleiche Sudokus mit vertauschten Zahlen
- ▶ Gleicher Schwierigkeitsgrad
- ▶ Unterschiedliche Featurevektoren bei gleichem Lösungsweg
- ▶ Lösung?



- ▶ Fast gleiche Sudokus mit vertauschten Zahlen
- ▶ Gleicher Schwierigkeitsgrad
- ▶ Unterschiedliche Featurevektoren bei gleichem Lösungsweg
- ▶ Sortierung der Features nach Häufigkeit
- ▶ Kein relevanter Informationsverlust
- ▶ Gleicher Featurevector auch bei vertauschten Ziffern

Entkopplung von konkreten Zahlen (Beispiel)

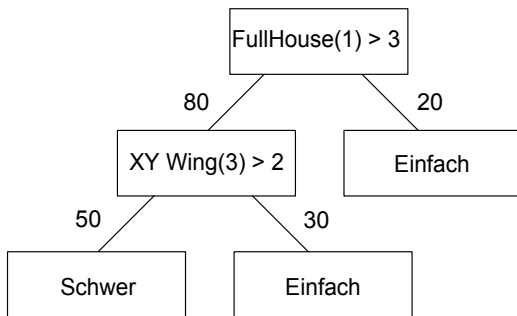


TECHNISCHE
UNIVERSITÄT
DARMSTADT

- ▶ Beispiel des Featurevectors einer Methode
 $(1, 0, 4, 15, 3, 0, 9, 2, 0)^T$
- ▶ Vertauschte Ziffern 7 und 8
 $(1, 0, 4, 15, 3, 0, 2, 9, 0)^T$
- ▶ Nach der Sortierung nach der Häufigkeit
 $(15, 9, 4, 3, 2, 1, 0, 0, 0)^T$



- ▶ Genauer Algorithmus, der auch Einblicke in die Klassifikationsgrundlage liefert
- ▶ Erstellt mit den Trainingsdaten einen Entscheidungsbaum





- ▶ Fremdsoftware für Klassifizierer und Lösungsmethoden
- ▶ Für den Klassifizierer: Weka¹, J48 Klassifizierer
- ▶ Für die Lösungsmethoden: Hodoku²
- ▶ Beide Projekte stehen unter der GPLv3 Lizenz
- ▶ Eigene Software in Java
- ▶ Extrahierung der Featurevektoren und Verbindung der Projekte

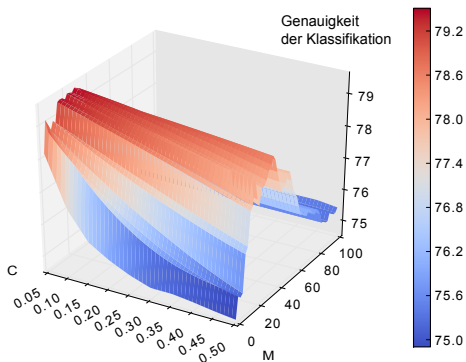
¹<http://www.cs.waikato.ac.nz/ml/weka/>

²<http://hodoku.sourceforge.net/de/index.php>



- ▶ J48 erhält 2 Parameter C und M
- ▶ Je kleiner C, desto eher wird der Baum abgeschnitten
- ▶ Je größer M, desto mehr Instanzen müssen sich mindestens in einem Blatt befinden
- ▶ Optimierung der Parameter für das verwendete Testset
- ▶ 1000 Auswertungen von Parameterkombinationen
- ▶ Optimum bei C: 0.1 und M: 30

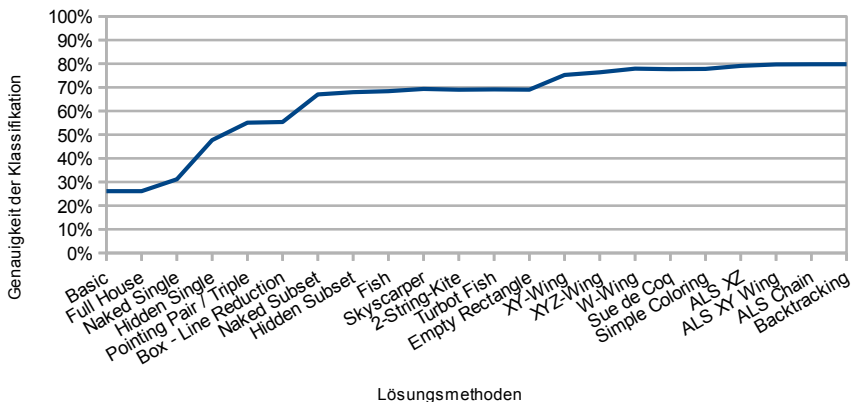
Optimierung der Parameter



Evaluierung des Featurevectors



TECHNISCHE
UNIVERSITÄT
DARMSTADT

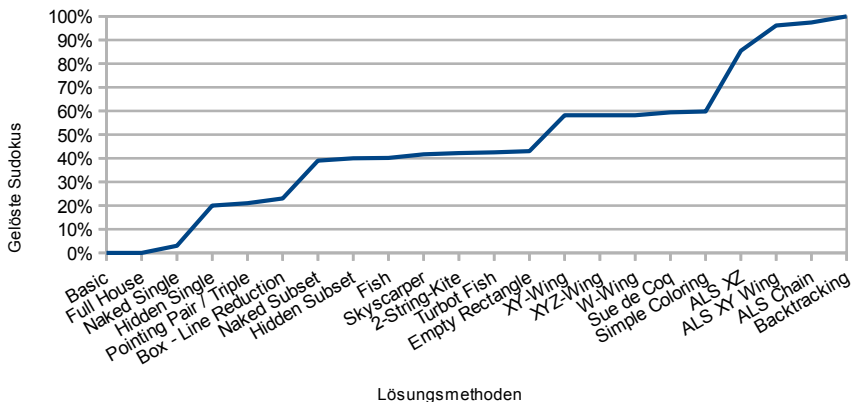


| | | Zugeordnete Klasse | | | | | Klasse |
|---|------|--------------------|-----|-----|-----|---|--------|
| | | a | b | c | d | e | |
| a | 1000 | 0 | 0 | 0 | 0 | 0 | |
| b | 0 | 568 | 115 | 93 | 224 | 0 | |
| c | 0 | 288 | 333 | 221 | 158 | 0 | |
| d | 0 | 309 | 257 | 225 | 209 | 0 | |
| e | 0 | 415 | 149 | 177 | 259 | 0 | |

a = Easy
b = Middle
c = Hard
d = Unfair
e = Extreme

Abbildung: Konfusionsmatrix mit einschließlich *Hidden Single* Methode

Evaluierung des Featurevectors



| | | Zugeordnete Klasse | | | | | Klasse |
|---|------|--------------------|-----|-----|-----|---|--------|
| | | a | b | c | d | e | |
| a | 1000 | 0 | 0 | 0 | 0 | 0 | |
| b | 0 | 999 | 1 | 0 | 0 | 0 | |
| c | 1 | 35 | 785 | 145 | 34 | 0 | |
| d | 0 | 10 | 129 | 636 | 225 | 0 | |
| e | 0 | 0 | 14 | 367 | 619 | 0 | |

a = Easy
b = Middle
c = Hard
d = Unfair
e = Extreme

Abbildung: Konfusionsmatrix mit allen vorgestellten Lösungsmethoden



- ▶ Vergleich verschiedener Bewertungsskalen
- ▶ Trainieren eines Klassifizierers mit einem Trainingsset
- ▶ Klassifizieren eines zweiten Sets mit dem gelernten Entscheidungsbaum
- ▶ Zur Verifikation Datensets vertauschen
- ▶ Darstellung der Ergebnisse als Matrix



| Ursprüngliche Klasse | Zugeordnete Klasse | | | | | |
|----------------------|--------------------|------|--------|------|--------|---------|
| | | Easy | Middle | Hard | Unfair | Extreme |
| | Sehr Einfach | 32 | 0 | 0 | 0 | 0 |
| | Einfach | 32 | 0 | 0 | 0 | 0 |
| | Standard | 32 | 0 | 0 | 0 | 0 |
| | Moderat | 0 | 32 | 0 | 0 | 0 |
| | Anspruchsvoll | 0 | 1 | 30 | 1 | 0 |
| | Sehr Anspruchsvoll | 0 | 1 | 28 | 2 | 1 |
| | Teuflisch | 0 | 0 | 7 | 8 | 17 |

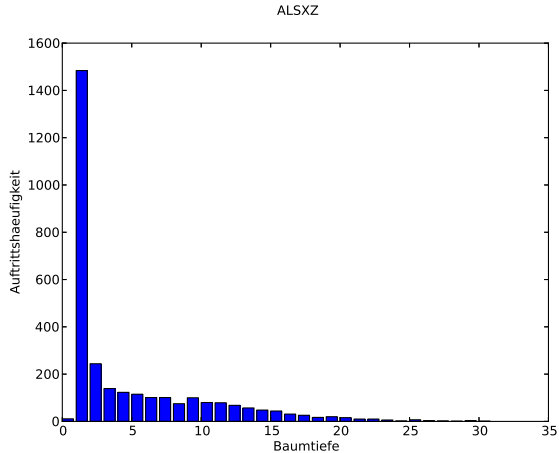


- ▶ Manche Features bieten mehr Informationen als andere
- ▶ Wichtigere Features stehen weiter oben im Entscheidungsbaum
- ▶ Unwichtige Features können auch gar nicht im Baum vorkommen
- ▶ J48 benutzt postpruning
- ▶ Evaluierung der wichtigen Features kann den Featurevector schmäler machen und dadurch die Laufzeit beim Erstellen des Entscheidungsbaums verbessern
- ▶ Finden der wichtigen Features mit CFS
- ▶ 18 von 261 Features ausgewählt, Genauigkeitsverlust von 1%

Relative Güte der Features (ALS XZ)



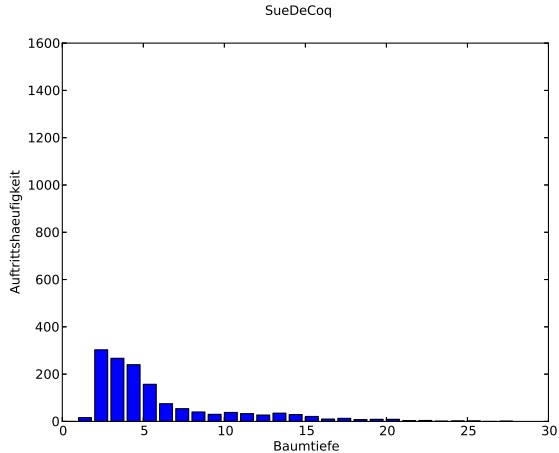
TECHNISCHE
UNIVERSITÄT
DARMSTADT



Relative Güte der Features (Sue de Coq)



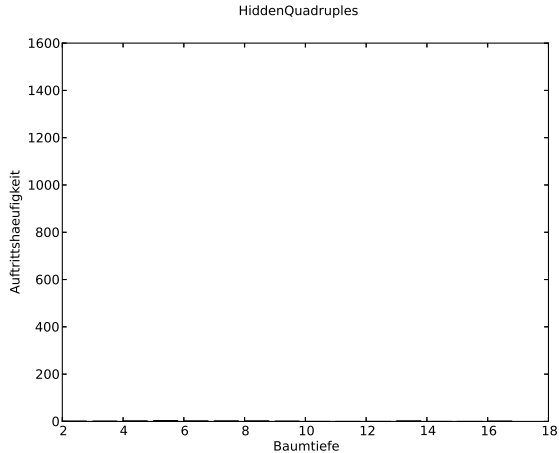
TECHNISCHE
UNIVERSITÄT
DARMSTADT



Relative Güte der Features (Hidden Quadruples)



TECHNISCHE
UNIVERSITÄT
DARMSTADT









- ▶ Erstellen eines Featurevectors
- ▶ Entkopplung von konkreten Zahlen
- ▶ Klassifikation mit J48
- ▶ Genauigkeit von ca. 80%
- ▶ Optimierung der Parameter
- ▶ Identifikation von Features die mehr Informationsgewinn liefern
- ▶ Mapping verschiedener Bewertungsskalen



- ▶ Entwicklung neuer Features z.B. Zeit, die ein menschlicher Spieler zum Lösen benötigt
- ▶ Verwendung von mehr Lösungsmethoden
- ▶ Evaluierung anderer Klassifikationsalgorithmen
- ▶ Qualitätsevaluation des Mappingverfahrens



-  Hall, M., A.: „Correlation-based Feature Subset Selection for Machine Learning “. University of Waikato, 1998.
-  Witten, I., H.; Frank, E.; Hall, M., A.: „Data Mining: Practical Machine Learning Tools and Techniques - Practical Machine Learning Tools and Techniques “. 3. Auflage, Elsevier Amsterdam, 2011.
-  Hobinger, B.: Hodoku Projekt
<http://hodoku.sourceforge.net/de/index.php>, retrieved on Apr 16, 2014
-  Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Ende...



TECHNISCHE
UNIVERSITÄT
DARMSTADT

