Lars Wohlrab

Studienarbeit

»Untersuchung verschiedener Strategien zur Behandlung unbekannter Attributwerte im SeCo-Regellerner«

Übersicht

- **Einleitung**
- Behandlungsstrategien
- Evaluation
- Schlussfolgerungen

Motivation

- Unbekannte Attributwerte
  - relevant für die meisten ML-Szenarien aus der Praxis
  - Heterogene Natur – verschiedene Ursachen und Semantiken
    - Unbekannter Wert ≠ Informationsverlust
- Tatsächliche Semantik oft unbekannt
  - Lerner kennen meist nur 1 Art von unbekannten Werten

Ansätze

- Mögliche Herangehensweisen:
  - Keine Beispiele mit unbekannten Werten
  - Keine Tests auf Attribute mit unbekannten Werten zulassen
  - Ersetzung von unbekannten Attributwerten durch „reguläre“
  - Einführung einer speziellen Semantik für unbekannte Werte

Zielsetzung

- Implementierung
  - Verschiedener Strategien zur Behandlung von unbekannten Attributwerten
  - Integriert: Ansatz zur Berücksichtigung von numerischer Unschärfe
- Evaluierung
- Basierend auf dem SeCo-Regellerner

Übersicht

- Einleitung
- **Behandlungsstrategien**
- Evaluation
- Schlussfolgerungen

4-Phasen-Modell

- An welchen Stellen muss man den Lerner erweitern?
- Umgang mit unbekannten Werten in drei Phasen:
  - Bewertung von Kandidatenregeln
  - Abtrennung der abgedeckten Trainingsbeispiele
  - Klassifikation neuer Beispiele
- …und für integrierte Umsetzung:
  - Vorverarbeitungsphase

Strategien I

- **Delete**
  - Entfernung aller Beispiele mit unbekannten Werten
  - Kann praktisch nur als Maßstab für minimal erreichbare Genauigkeit dienen
  - Problem: Verschwendung von Trainingsinformation
- **Ignore**
  - Unbekannte Werte werden nie abgedeckt
  - Keine Verschwendung von Information
  - Lernen und Klassifizieren nur auf Grundlage bekannter Angaben

Strategien II

- **AnyValue**
  - Unbekannte Werte werden von jeder Bedingung abgedeckt – optimistisches Gegenstück zu Ignore
  - Führt zu geringerer „Selektivität“ von Bedingungen
    - Schlechteres Lernen aus unvollständigen Attributen
    - Größere Modelle
- **SpecialValue**
  - „unbekannt“ wird als eigenständiger Attributwert behandelt
  - Ignore + Option, aus dem Fehlen von Werten zu lernen

Strategien III

- **Common**
  - Ersetzung unbekannter Werte durch häufigsten Wert/Mittelwert
  - Minimiert den Ersetzungsfehler
- **NN**
  - Verbesserung der Qualität der eingesetzten Schätzer
  - Durch „Lokalisierung“ der Schätzungen
- **DBI** – verteilungsbasierte Ersetzung
  - Berücksichtigung der Verteilung des Attributs
    - Einsetzung aller möglichen Werte, gewichtet mit Wkt.
    - Begrenzung der Aufspaltung durch Mindestgewicht
    - Partielle Abdeckung von Beispielen möglich
  - Wann ersetzt man?

Strategien IV

- **HP – Heuristic Penalty**
  - „Bestrafung“ der Bewertungsheuristik für Tests auf unbekannte Werte
    - Von Entscheidungsbäumen bekannt als Prinzip des „reduced information gain“
  - Tests auf unbekannte Werte werden immer als Fehler gezählt
  - Integriert: Umgang mit numerischer Unschärfe
    - Parametrisiert mit Unschärferadius
    - Beispiel kann nicht abgedeckt werden von Test auf Wert im Unschärfebereich

Übersicht

- Einleitung
- Behandlungsstrategien
- **Evaluation**
- Schlussfolgerungen

Datensätze

- …mit unbekannten Werten
  - Reale Daten mit „Lücken“
  - Manuelles „Ausdünnen“ – präparierte Daten

  Pro:
  - Vergleich mit dem Ergebnis ohne unbekannte Werte möglich
  - Kontrolle über betroffene Attribute und Ausfallraten

  Contra:
  - Systematischer Einfluss des Erzeugungsverfahrens

Präparierte Daten

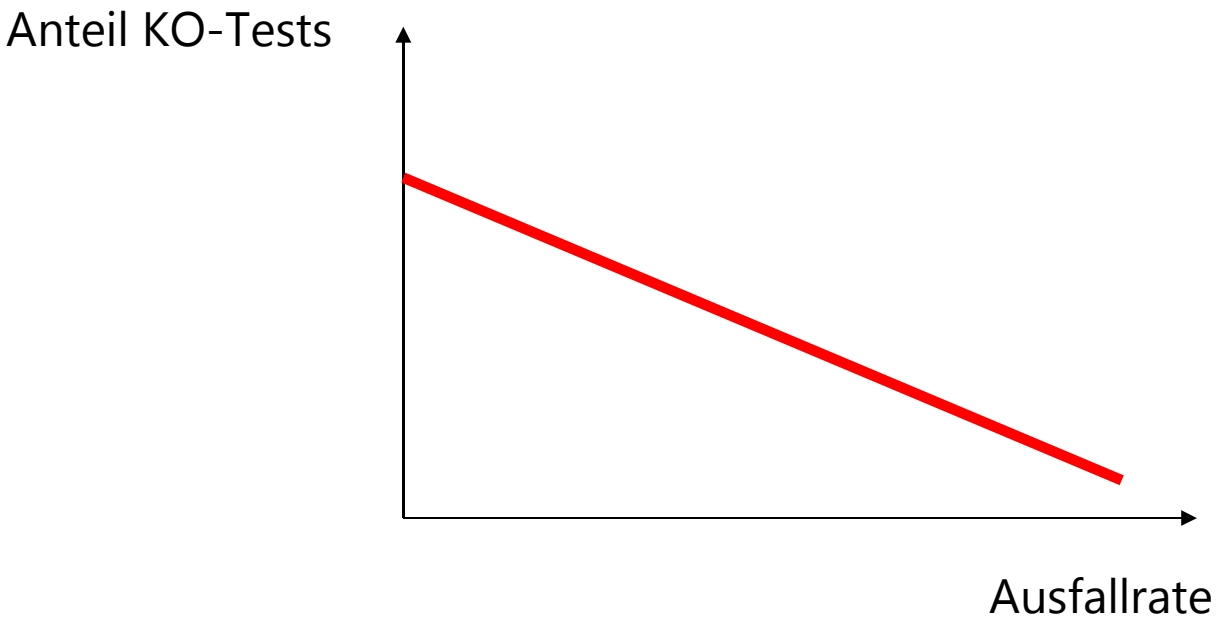
- Erzeugungsalgorithmus
  - Auswahl der 3 wertvollsten Attribute
  - Entfernung von zufälligen Werten unabhängig voneinander in 15%-Schritten bis max. 90%

- Mit den 3 umfangreichsten Datensätzen ohne unbekannte Werte

Ergebnisse

Präparierte Daten

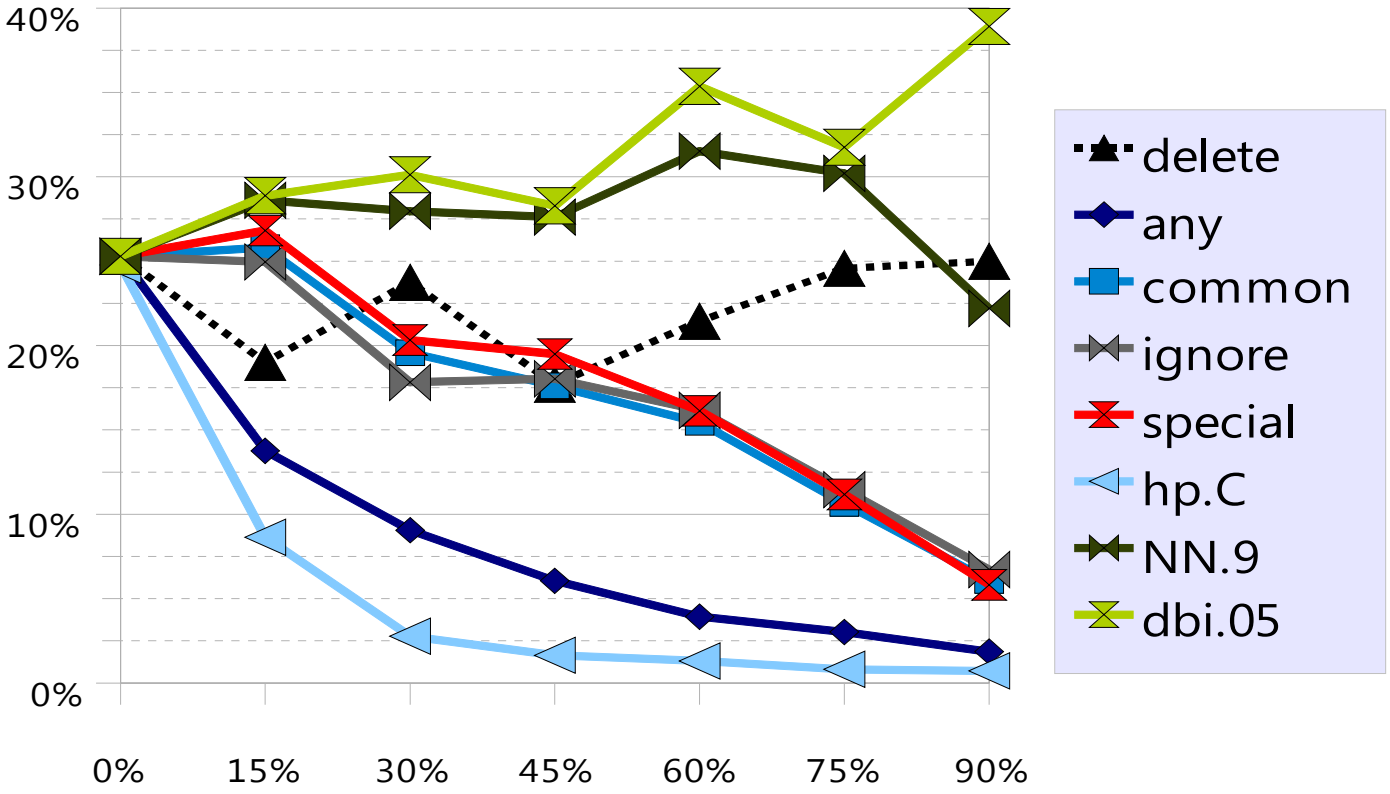
- Einfluss auf gelernte Modelle - **Erwartung**



Ergebnisse

Präparierte Daten

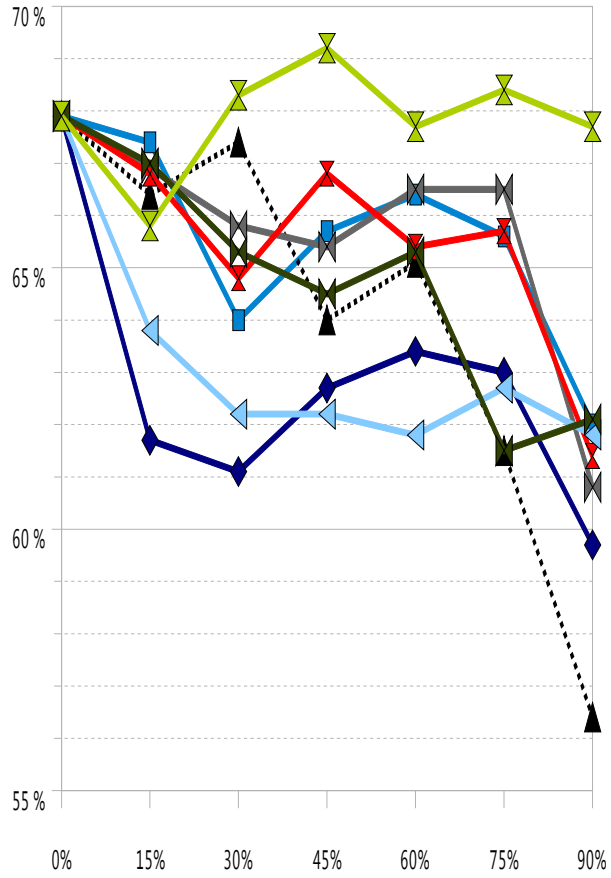
- Einfluss auf gelernte Modelle

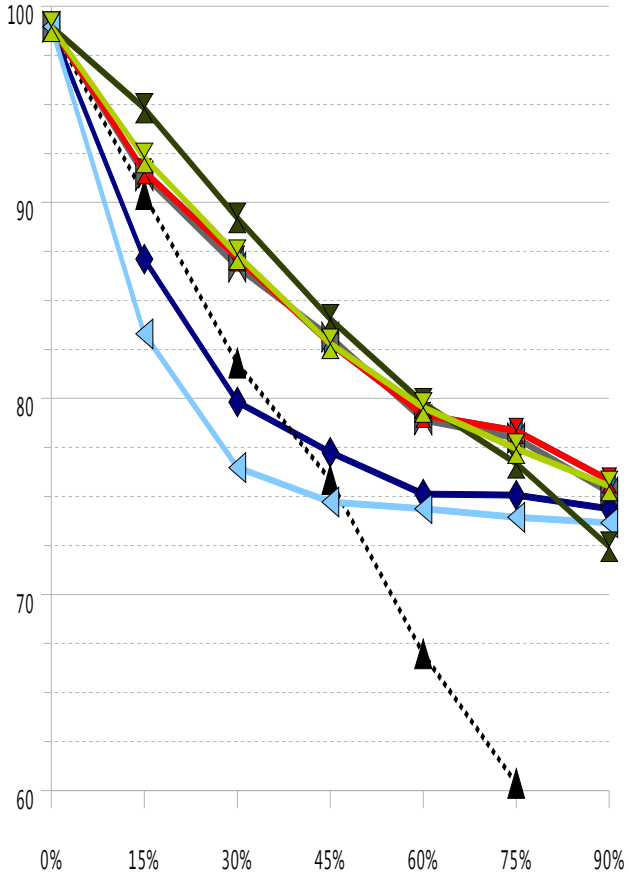


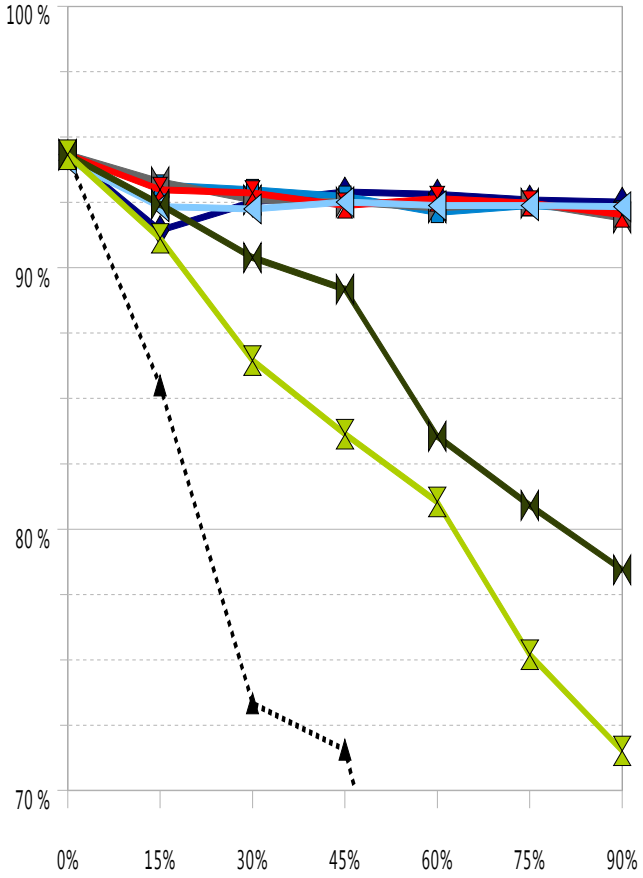
Ergebnisse

Präparierte Daten

- Erzielte Genauigkeiten – präparierte Daten



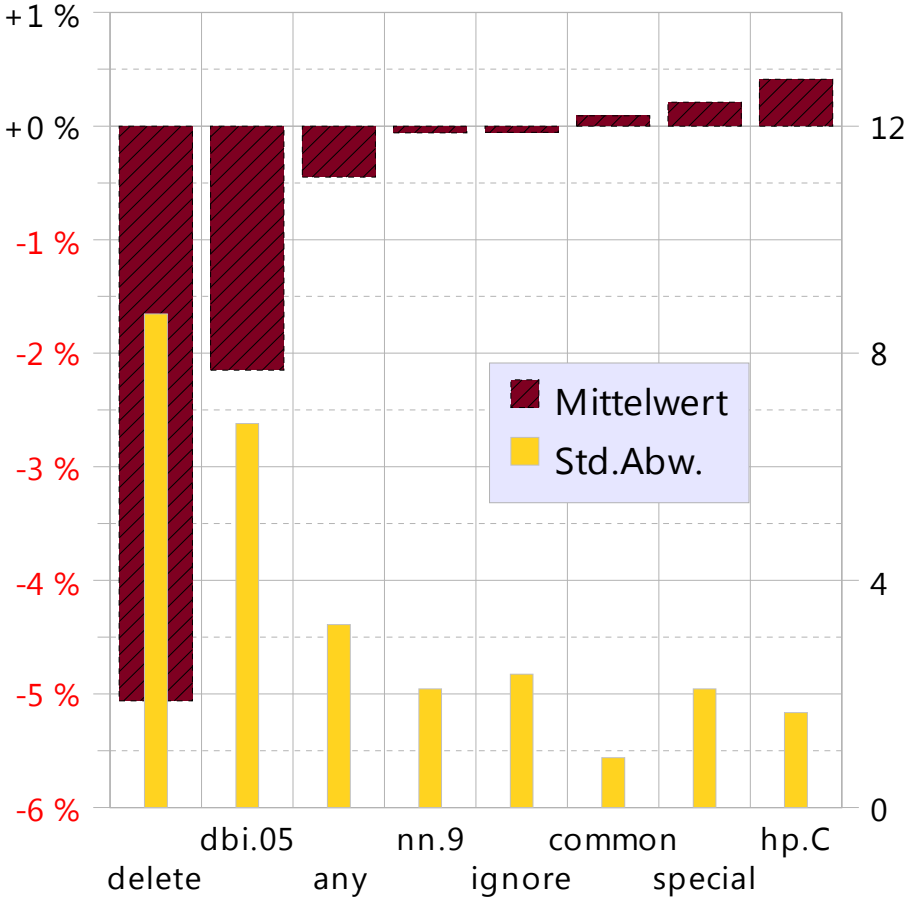




Ergebnisse

Reale Daten

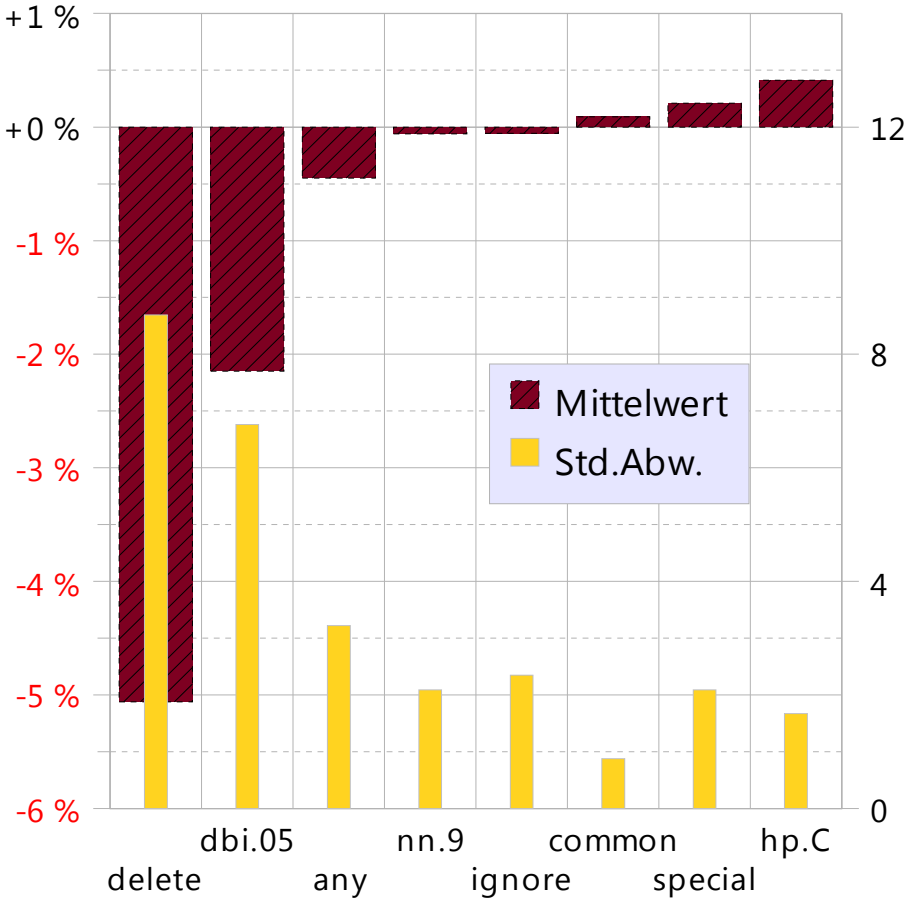
- Mittlere Genauigkeiten – reale Daten

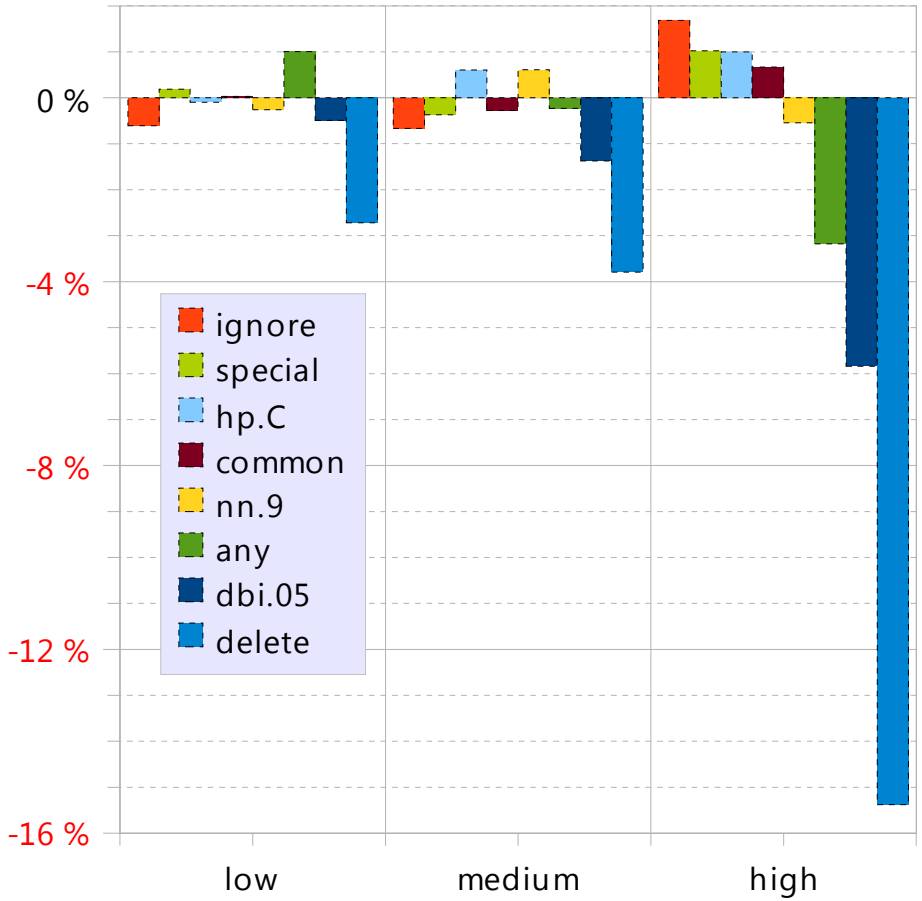


Ergebnisse

Reale Daten

- Mittlere Genauigkeiten – reale Daten





Ergebnisse

Reale Daten

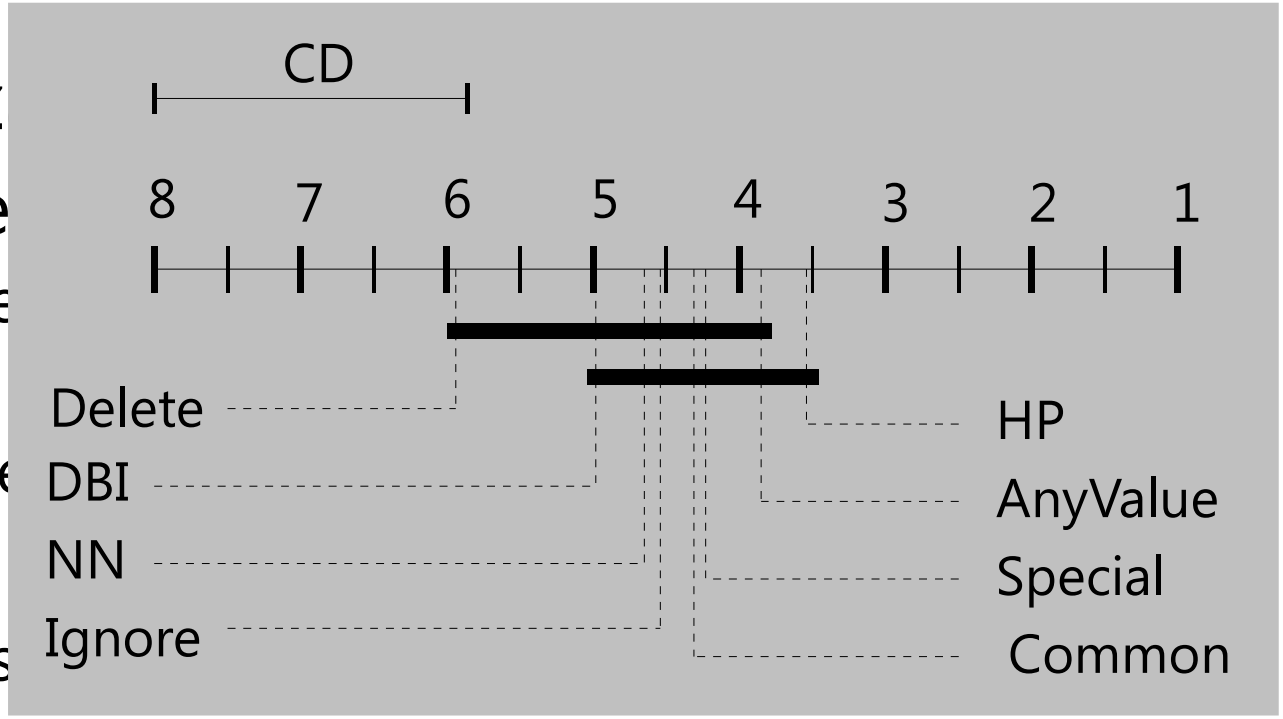
- Signifikanzaussagen
  - paarweise t-Tests
    - (fast) alle Strategien sind signifikant besser als Delete
    - Alle anderen Unterschiede sind zu gering
  - Rank-Test (Friedman/Nemenyi)
    - Nur HP sicher besser als Delete

	Delete	DBI	NN	Ignore	Common	Special	Any	HP
# winner	2	4	3	4	3	6	4	7
Ø rank	5,9	5,0	4,6	4,5	4,3	4,2	3,9	3,5

Ergebnisse

Reale Daten

- Signifikanz
  - paarweise
    - (fast) alle Delete
    - Alle ande
  - Rank-Test
    - Nur HP s



	Delete	DBI	NN	Ignore	Common	Special	Any	HP
# winner	2	4	3	4	3	6	4	7
Ø rank	5,9	5,0	4,6	4,5	4,3	4,2	3,9	3,5

Ergebnisse

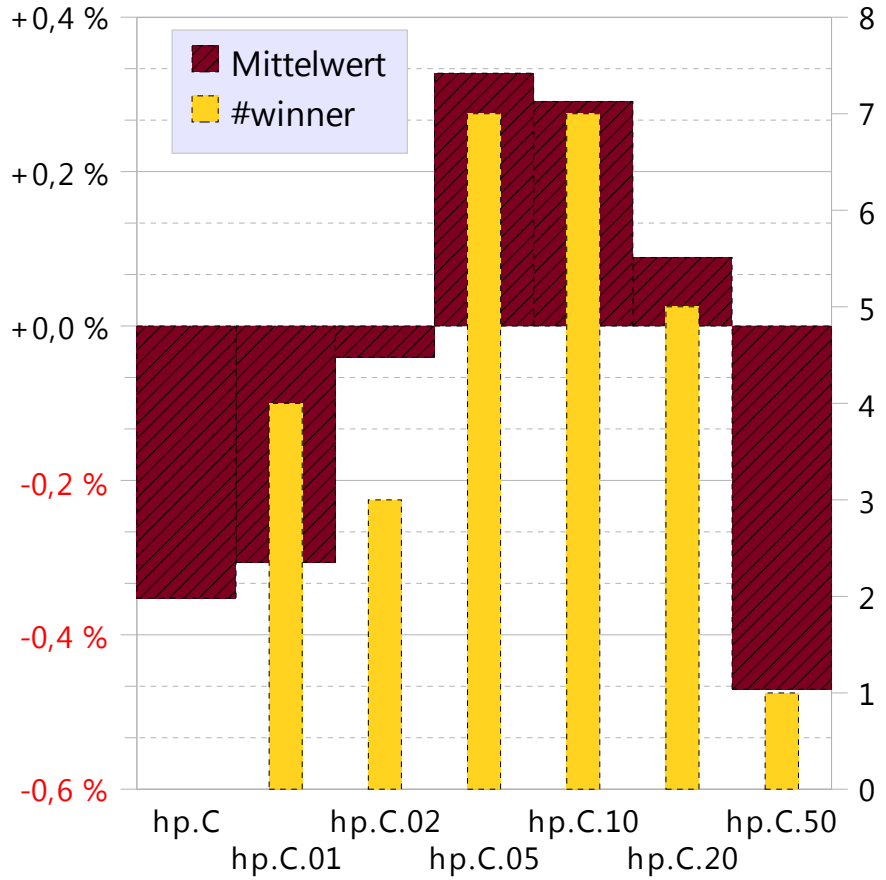
Reale Daten

- HP mit Numerischer Unschärfe (NUS)
  - Nur rudimentäre Untersuchung
  - ohne Domänenwissen, ohne Rücksicht auf Wertebereiche
  - gleiche Unschärfe-Intervalle für alle Attribute
    - 6 feste Werte zwischen 0,01 und 0,5

Ergebnisse

Reale Daten

- im Mittel fast immer besser als die Basis-variante ohne NUS
  - Basisvariante auf keinem Datensatz optimal
  - Beste NUS-Schranke im Mittel 2%, maximal 8% besser als HP ohne NUS



Übersicht

- Einleitung
- Behandlungsstrategien
- Evaluation
- **Schlussfolgerungen**

Schlussfolgerungen I

- Relativ starker Einfluss der Datensätze
  - Keine der Strategien gleichermaßen für alle Datensätze geeignet
  - Auch im Mittel schwächere Strategien auf einzelnen Datensätzen deutlich überlegen
  - Nur Delete eindeutig suboptimal

→ flexible Wahl der anzuwendenden Strategie ist für einen Lerner von Vorteil

Schlussfolgerungen II

- HP-Strategie hinterlässt zwiespältigem Eindruck
  - Explizite Bestrafung der Bewertungsheuristik offenbar nicht qualitätssensitiv
  - Sehr gute Resultate auf realen Daten
  - integrierte NUS-Unterstützung durchaus vielversprechend
    - attributspezifische Unschärfeschranken
    - auch unabhängig vom HP-Ansatz

Vielen Dank