Marcel Spranger

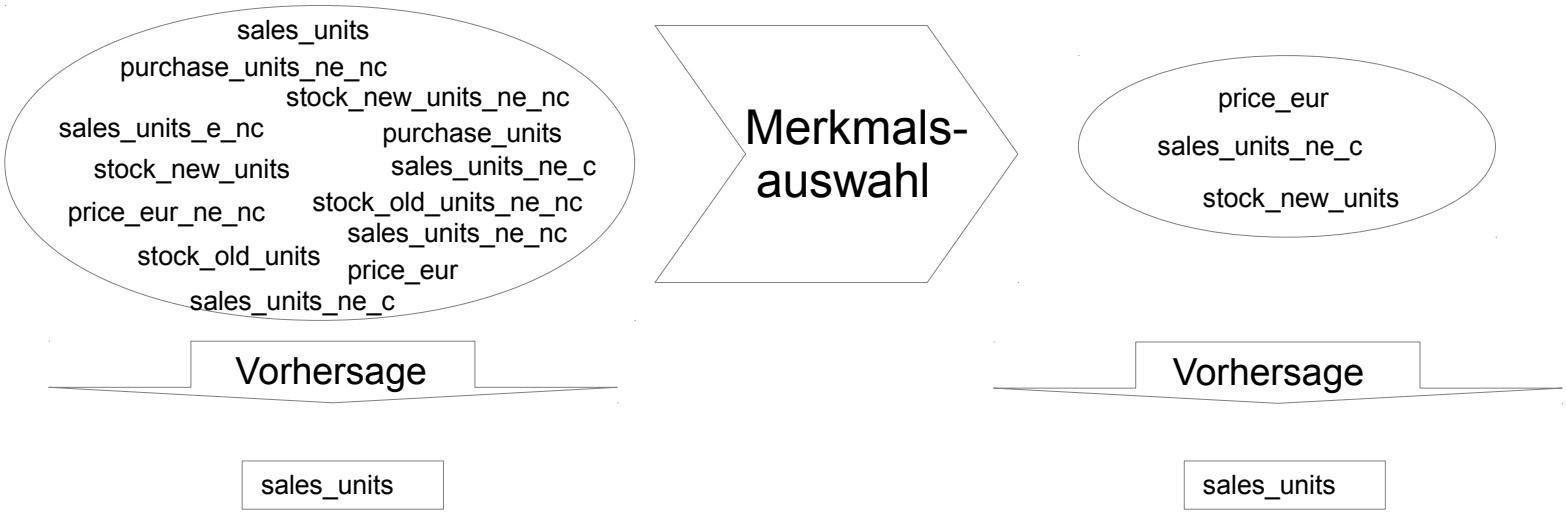
Merkmalsauswahl zur Optimierung von Prognoseprozessen auf Verkaufsdaten

15. Dezember 2014

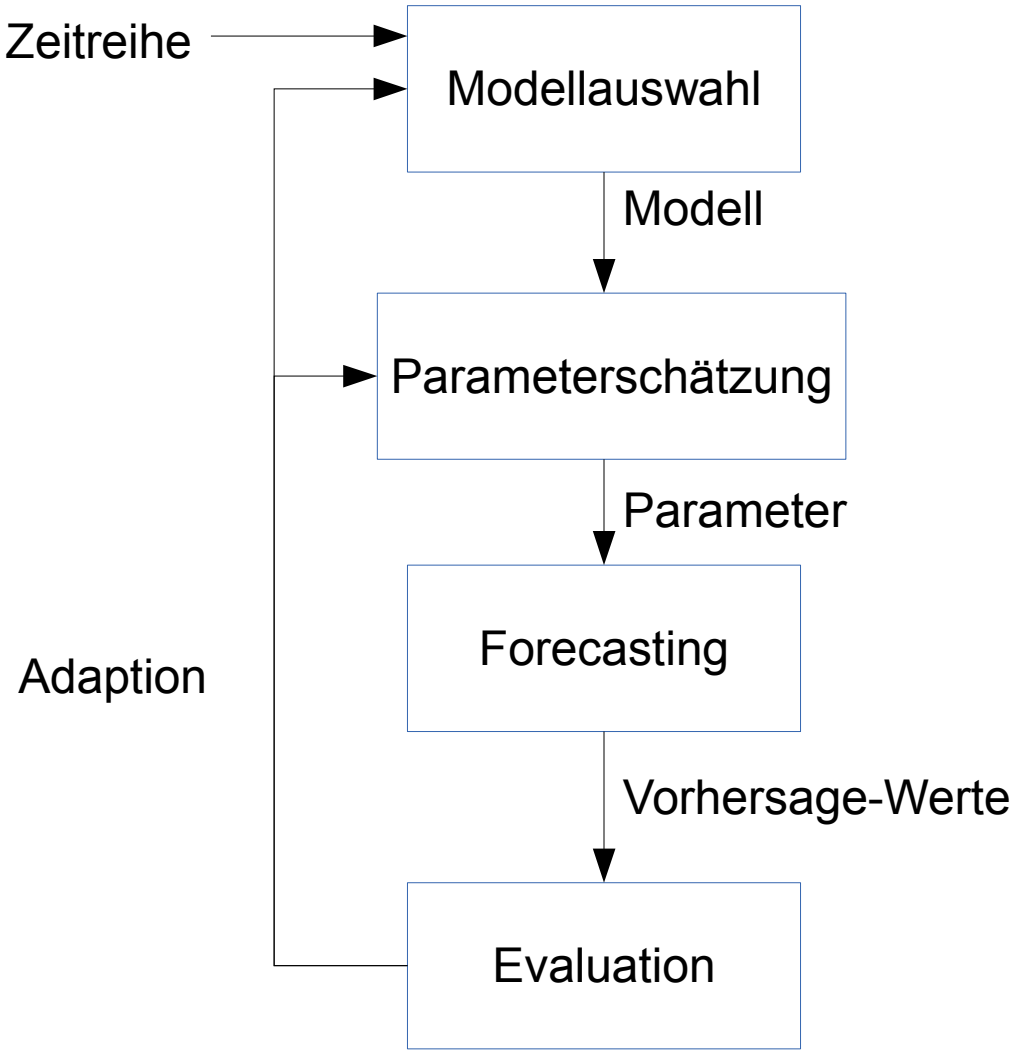
- Zeitreihenvorhersage: Prognose zukünftiger Werte einer Zeitreihe

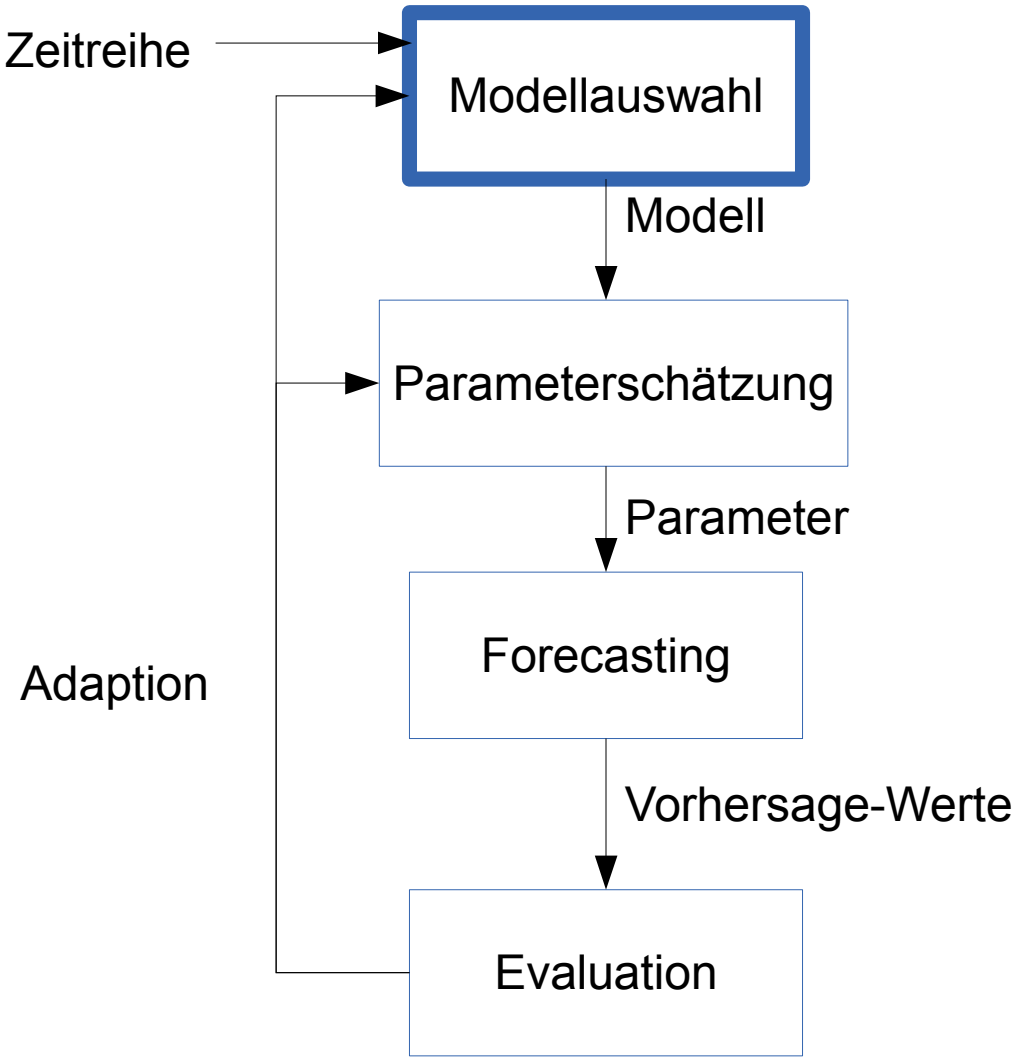
- Cross-Sectionale Forecasting: Prognose des Zielattributs anhand weiterer Attribute

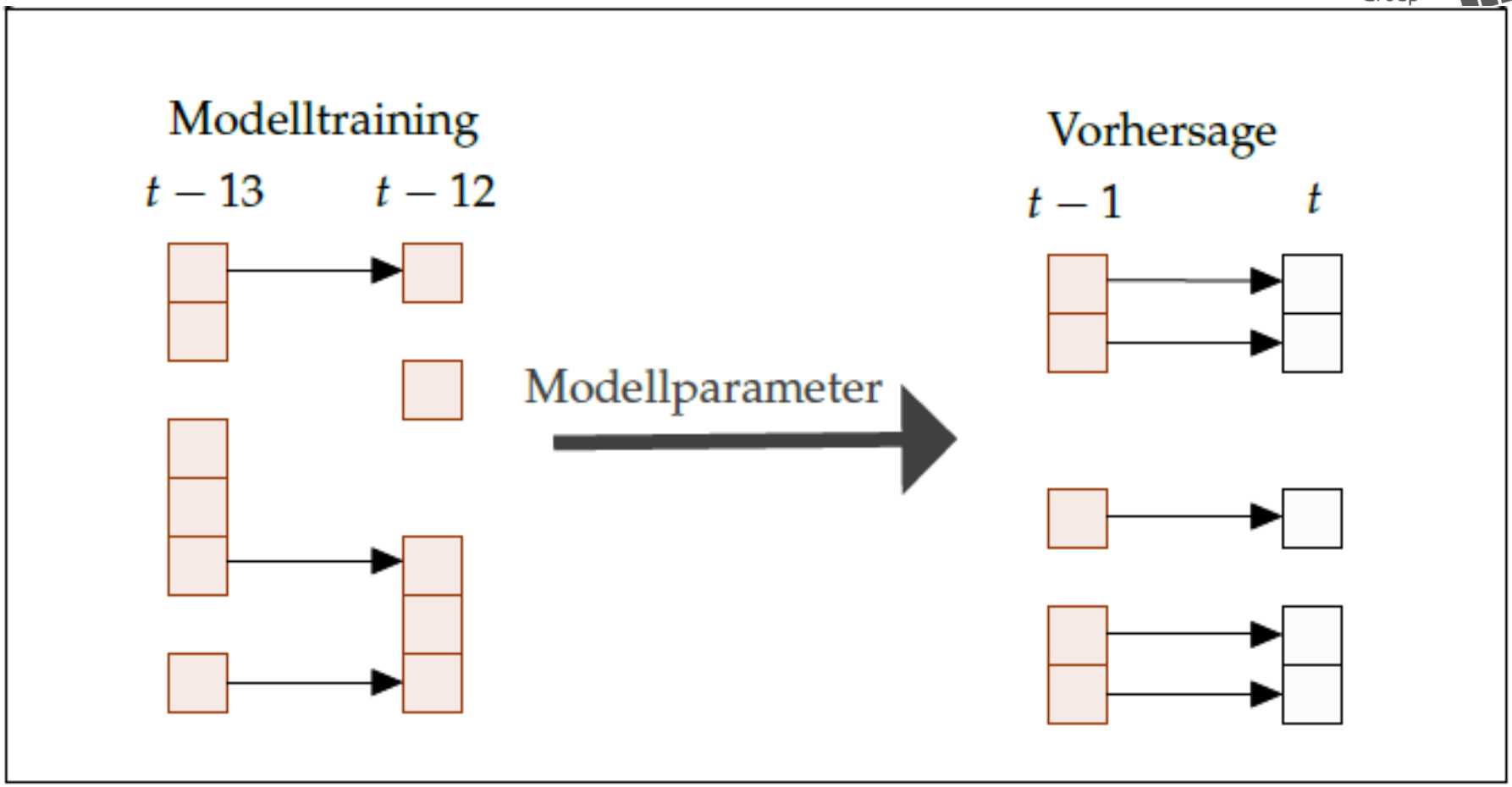
- Beispiele:
  - Vorhersage von Energieverbrauch und Energieerzeugung
  - Bruttosozialprodukt, Arbeitslosenquote
  - Verkaufszahlen



- Motivation
- Grundlagen
- Wrapper - Ansatz
- Filter - Ansatz
- Hybrider Ansatz
- Evaluation
- Zusammenfassung


Zeitreihe
Modellauswahl
Modell
Parameterschätzung
Parameter
Forecasting
Adaption
Vorhersage-Werte
Evaluation

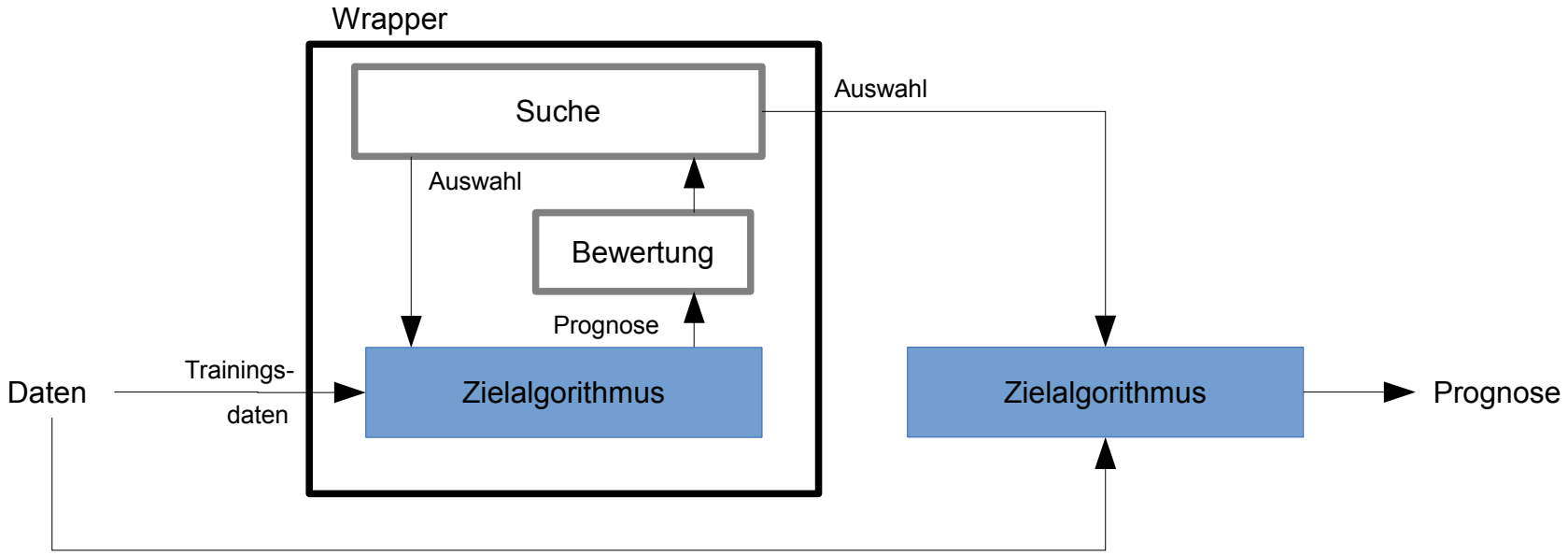

Zeitreihe
Modellauswahl
Modell
Parameterschätzung
Parameter
Forecasting
Adaption
Vorhersage-Werte
Evaluation



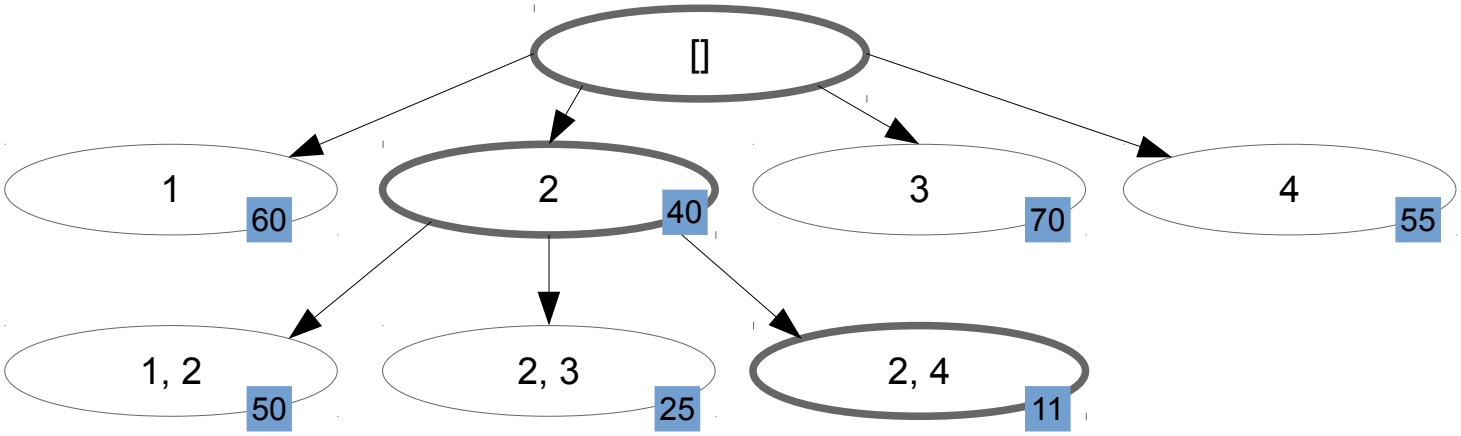
$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + u$$

- Motivation
- Grundlagen
- Wrapper - Ansatz
- Filter - Ansatz
- Hybrider Ansatz
- Evaluation
- Zusammenfassung

- Idee: Zielalgorithmus als Blackbox zur Bewertung der Auswahl
- Bewertung wird auf den Trainingsdaten ausgeführt
- Suche nach Attributskombination mit bester Bewertung
- Liefert im allgemeinen sehr gute Ergebnisse, da stark an den Zielalgorithmus angepasst
- Jedoch sehr aufwendig



Forward Selection & Backward Elimination



Forward Selection

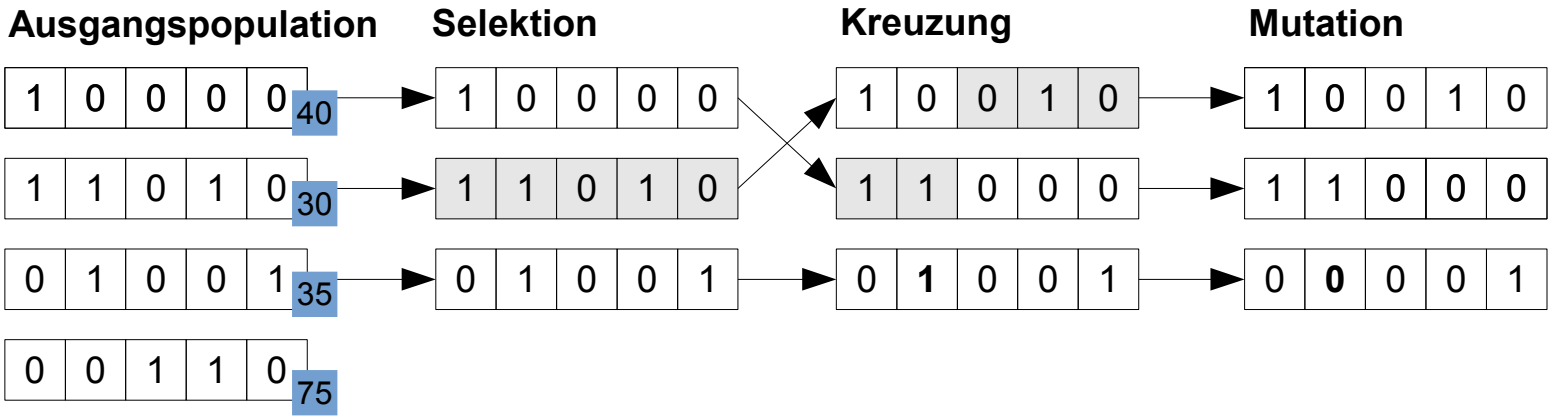
- Beginne mit leerer Attributskombination
- Überprüfe, welche Attributshinzunahme zur stärksten Verbesserung führt
- Füge so lange Attribute hinzu, bis keine Verbesserung mehr stattfindet

Backward Elimination

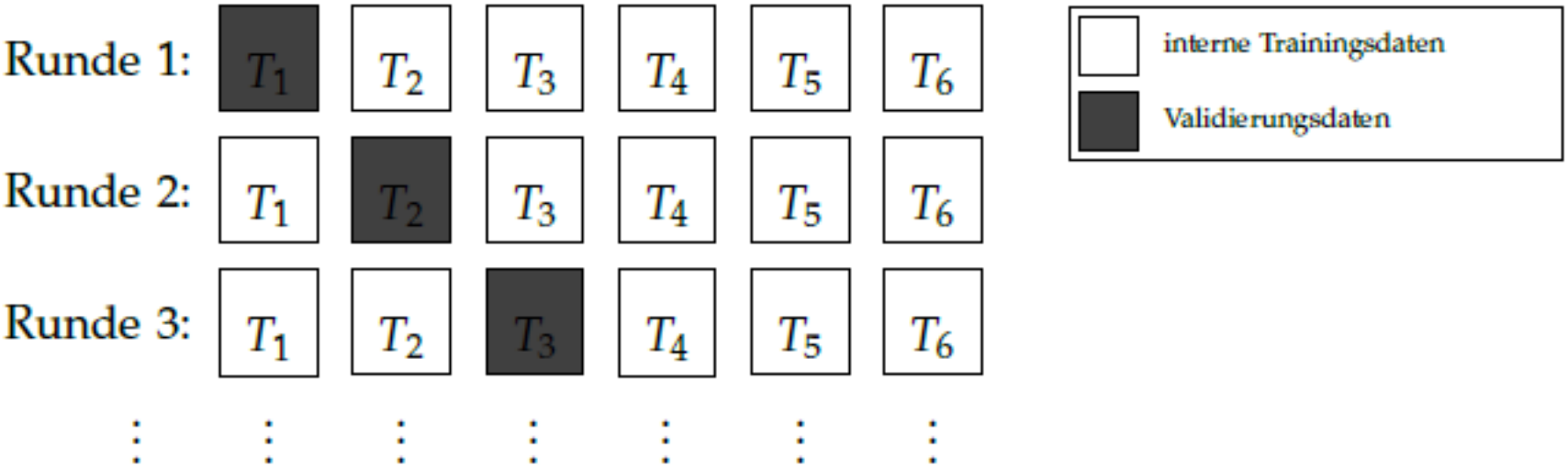
- Nehme alle Attribute in Kombination auf
- Überprüfe, welche Attributentfernung zur stärksten Verbesserung führt
- Entferne so lange, bis keine Verbesserung mehr stattfindet

Genetischer Algorithmus

- Parameterbelegung stellt ein Individuum dar
- Durch Kreuzen zweier Individuen und Mutationen entstehen neue Individuen
- Nur Individuen mit ausreichender Bewertung dürfen in der nächsten Iteration neue Nachfahren erzeugen

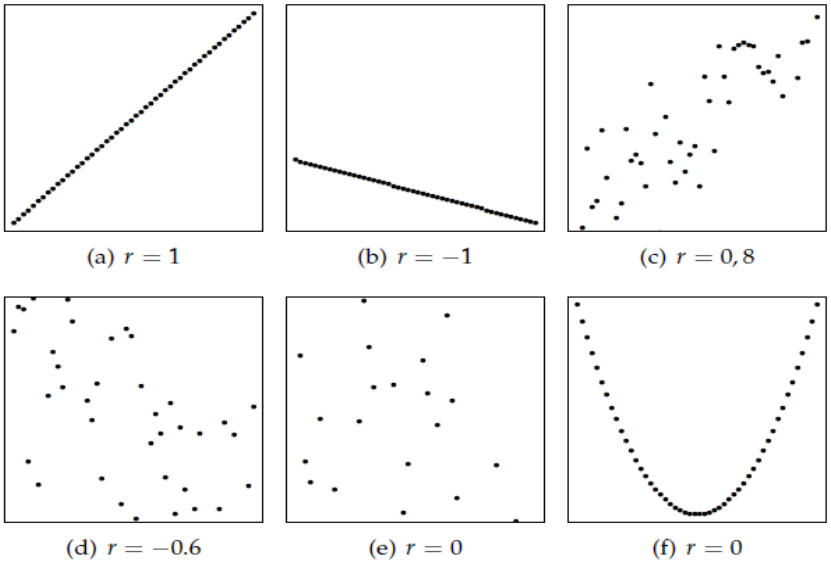


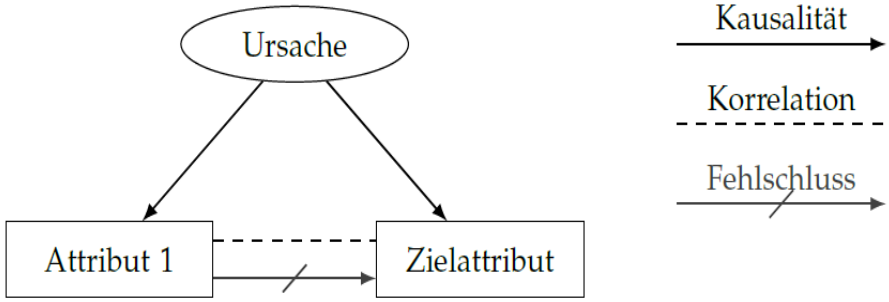
- Problem: Überfittung
- Trainingsdaten werden mehrmals in interne Trainingsdaten und Validierungsdaten aufgeteilt
- Als Rückgabe dient die durchschnittliche Bewertung

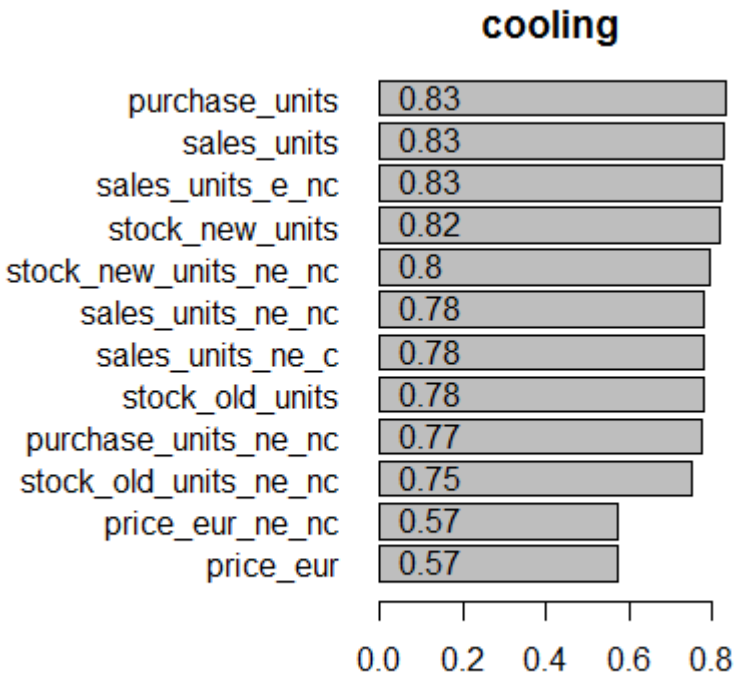


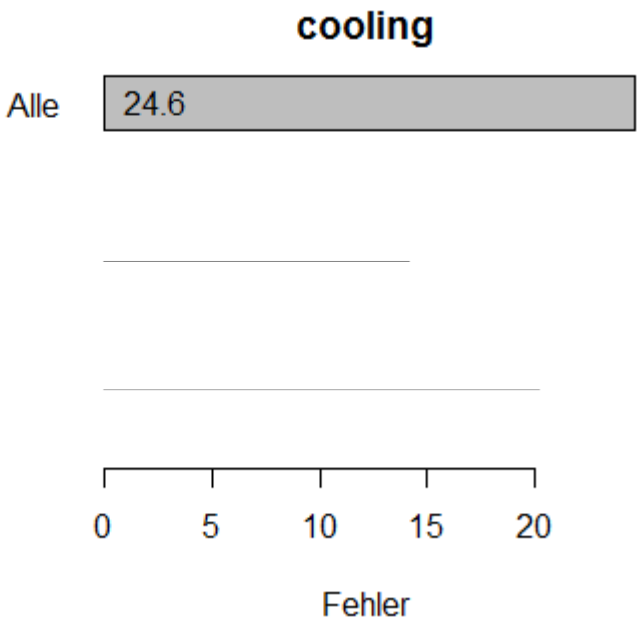
- Motivation
- Grundlagen
- Wrapper - Ansatz
- Filter - Ansatz
- Hybrider Ansatz
- Evaluation
- Zusammenfassung

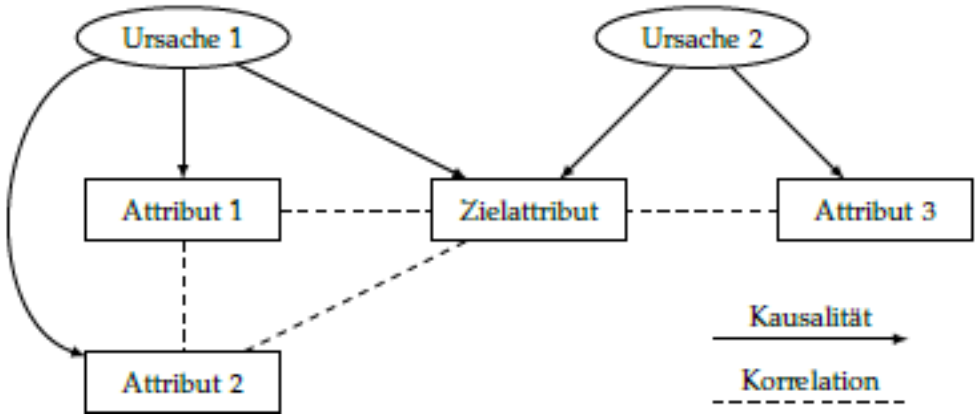
- Idee: die besten Merkmale werden anhand statistischer Eigenschaften herausgefiltert
- Unabhängig von später verwendeten Algorithmus, deswegen weniger genau als Wrapper-Methode
- Dafür weniger Aufwendig, daher schneller
- Korrelationskoeffizient nach Bravais und Pearson $$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

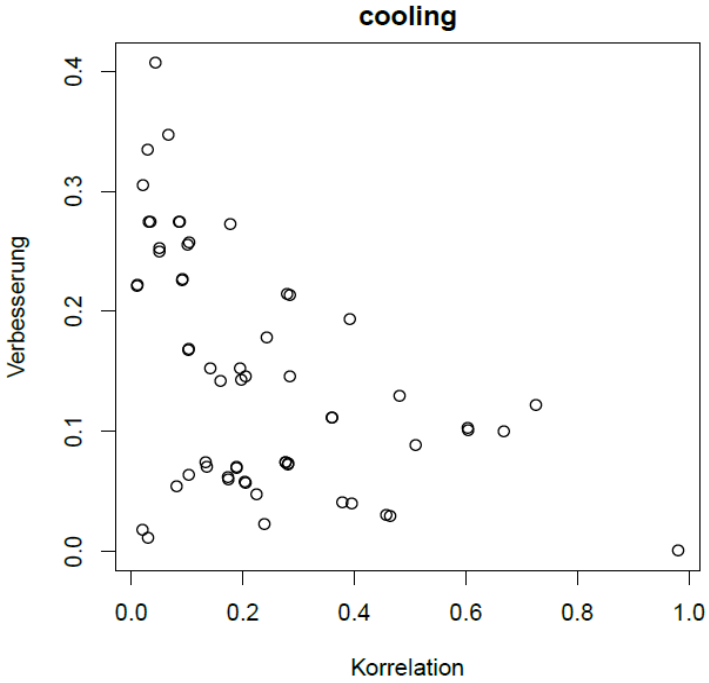





cooling
purchase_units 0.83
sales_units 0.83
sales_units_e_nc 0.83
stock_new_units 0.82
stock_new_units_ne_nc 0.8
sales_units_ne_nc 0.78
sales_units_ne_c 0.78
stock_old_units 0.78
purchase_units_ne_nc 0.77
stock_old_units_ne_nc 0.75
price_eur_ne_nc 0.57
price_eur 0.57
0.0 0.2 0.4 0.6 0.8
Korrelation


cooling
Alle 24.6
0 5 10 15 20
Fehler


Ursache 1
Ursache 2
Attribut 1
Zielattribut
Attribut 3
Attribut 2
Kausalität
Korrelation



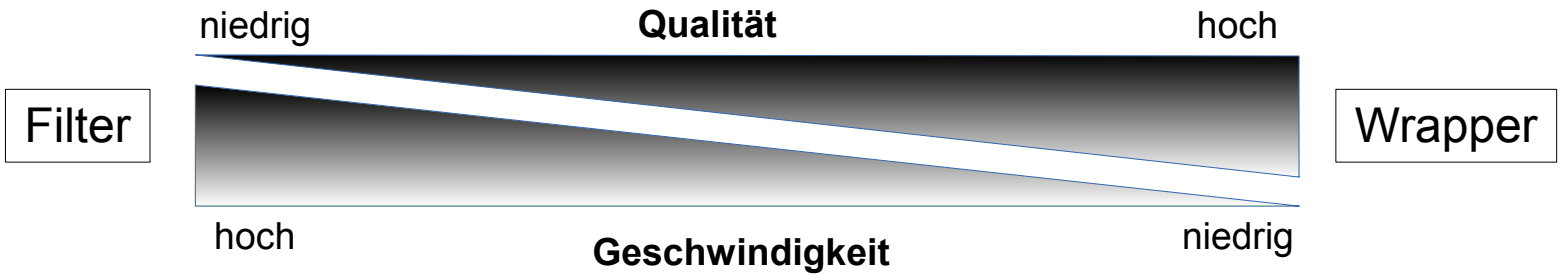
- strarke Korrelation → kaum Verbesserung
- Geringe Korrelation
  → starke Verbesserung möglich
- Aber: Geringe Korrelation kein Garant für Verbesserung!

- Korrelationsbasierte Filter nach Hall
- Merkmale sollen hohe Korrelation zu Zielattribut besitzen, jedoch geringe Korrelation untereinander

$$Merit_S = \frac{k\,\overline{r_{cf}}}{\sqrt{k + k(k-1)\,\overline{r_{ff}}}}$$

- Motivation
- Grundlagen
- Wrapper - Ansatz
- Filter - Ansatz
- Hybrider Ansatz
- Evaluation
- Zusammenfassung

- Filter-Ansatz schnell, aber ungenau
- Wrapper-Ansatz genau, aber langsam



→ Kombination beider Ansätze

Filter-Wrapper

Filter gibt nur bestimmte Anzahl von Kombinationen zurück, auf die der Wrapper-Ansatz durchgeführt wird.

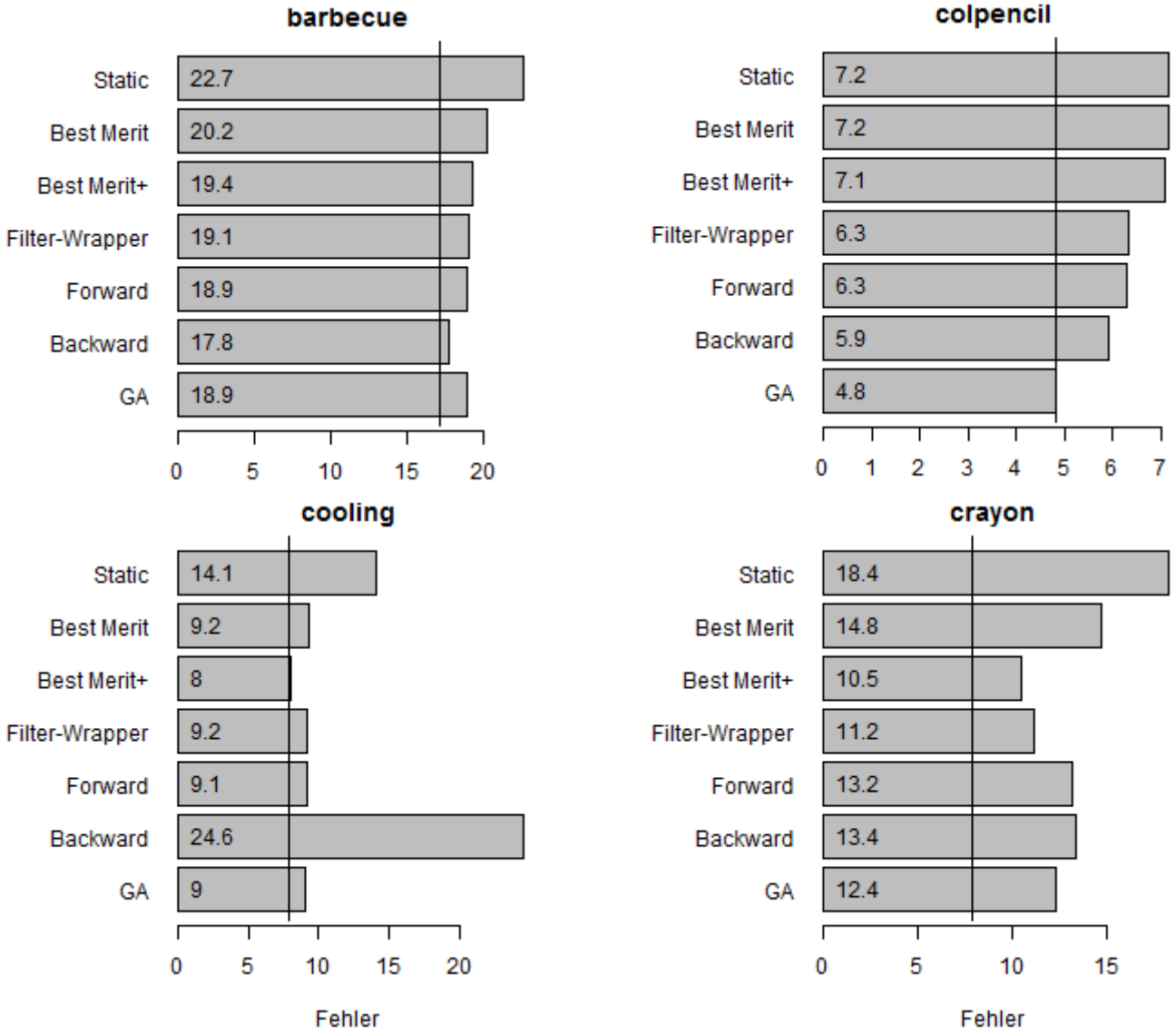
Merit+

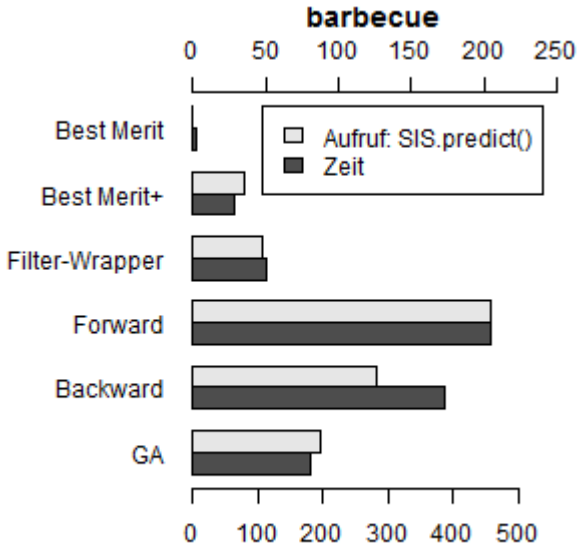
Wrapper-Ansatz zur Korrelationsbestimmung der Einzelattribute.

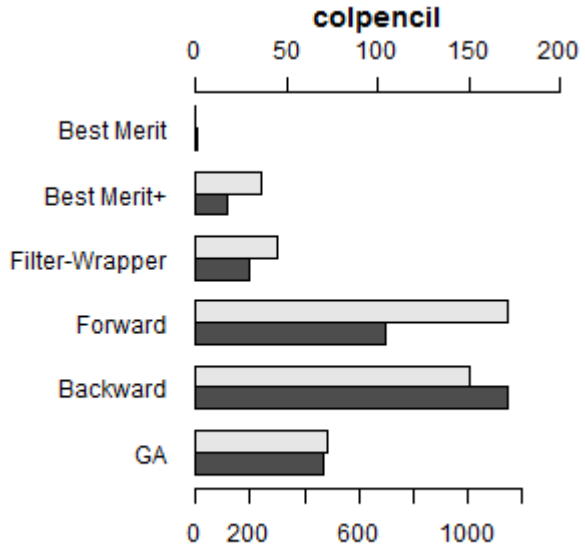
- Motivation
- Grundlagen
- Wrapper - Ansatz
- Filter - Ansatz
- Hybrider Ansatz
- Evaluation
- Zusammenfassung

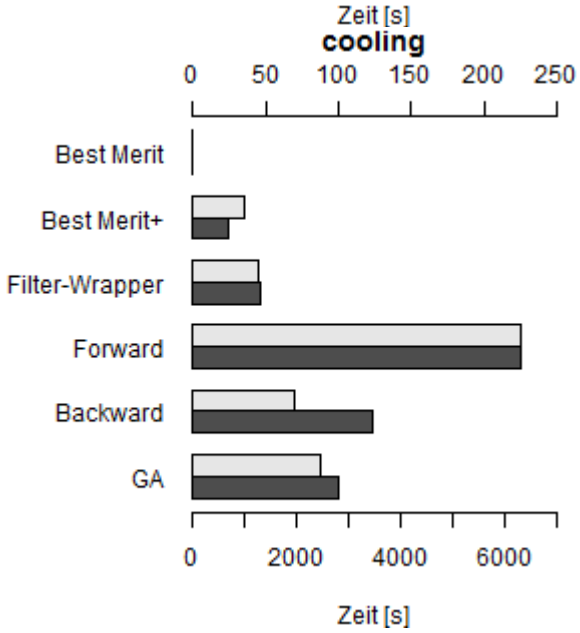
- Bestimmung der Merkmale mithilfe Filter-, Wrapper- oder Hybriden Ansatz anhand der Trainingsdaten
- Verwendung dieser Merkmale zur Vorhersage der Evaluationsdaten
- Bestimmung der Abweichung/ des Fehlers der entstandenen Vorhersagen
- Methoden:

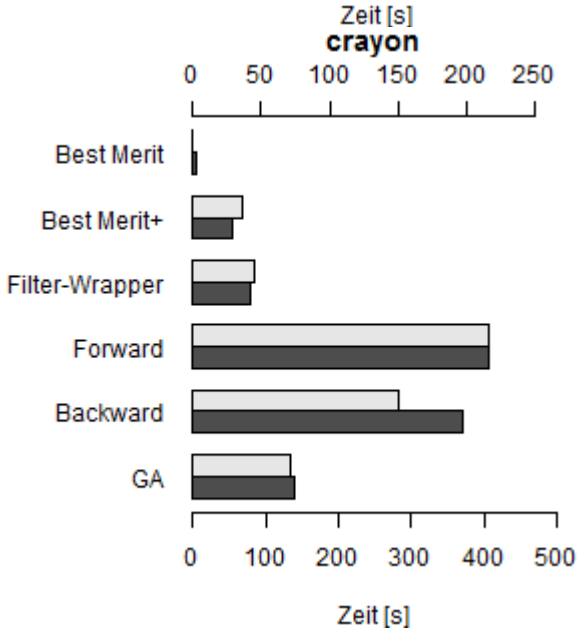
  - Static: Ziel-Attribut als Regressionsattribut (zum Vergleich)
  - Best Merit
  - Best Merit+
  - Filter-Wrapper: 15 besten Filter-Ergebnisse werden von Wrapper überprüft
  - Forward Selection
  - Backward Elimination
  - GA
  - Bestmögliches Ergebnis (zum Vergleich)


barbecue
Static
Best Merit
Best Merit+
Filter-Wrapper
Forward
Backward
GA
22.7
20.2
19.4
19.1
18.9
17.8
18.9
0
5
10
15
20
colpencil
Static
Best Merit
Best Merit+
Filter-Wrapper
Forward
Backward
GA
7.2
7.2
7.1
6.3
6.3
5.9
4.8
0
1
2
3
4
5
6
7
cooling
Static
Best Merit
Best Merit+
Filter-Wrapper
Forward
Backward
GA
14.1
9.2
8
9.2
9.1
24.6
9
0
5
10
15
20
Fehler
crayon
Static
Best Merit
Best Merit+
Filter-Wrapper
Forward
Backward
GA
18.4
14.8
10.5
11.2
13.2
13.4
12.4
0
5
10
15
Fehler


barbecue
0 50 100 150 200 250
Best Merit
Best Merit+
Filter-Wrapper
Forward
Backward
GA
Aufruf: SIS.predict()
Zeit
0 100 200 300 400 500
Zeit [s]


colpencil
0 50 100 150 200
Best Merit
Best Merit+
Filter-Wrapper
Forward
Backward
GA
0 200 600 1000
Zeit [s]


cooling
0 50 100 150 200 250
Best Merit
Best Merit+
Filter-Wrapper
Forward
Backward
GA
0 2000 4000 6000
Zeit [s]


crayon
0 50 100 150 200 250
Best Merit
Best Merit+
Filter-Wrapper
Forward
Backward
GA
0 100 200 300 400 500
Zeit [s]

- Wahl der Modellatribute hat entscheidenden Einfluss auf die Vorhersagequalität
- Automatisierte Merkmalsauswahl ist potentiell in der Lage, Qualität der Vorhersage erheblich zu Verbessern im Vergleich zur statischen Attributsauswahl
- Bis auf Ausnahmen lieferten sowohl Filter-, Wrapper- und kombinierte Ansätze gute bis sehr gute Ergebnisse
- Ein klarer Favorit konnte jedoch nicht bestimmt werden
- Future Work:
  - Nichtlineare Zusammenhänge
  - Hohe Attributanzahl
  - Wann ist welche Methode anzuwenden?

Marcel Spranger

Merkmalsauswahl zur Optimierung von Prognoseprozessen auf Verkaufsdaten

15. Dezember 2014