Dresden Database Systems Group

Verteidigung der Bachelorarbeit

MERKMALSAUSWAHL ZUR OPTIMIERUNG VON PROGNOSEPROZESSEN

Von:

Tom Fels

23.11.2015

Betreut durch: Prof. Dr.-Ing. Wolfgang Lehner

Dresden Database
Systems Group

Motivation

Motivation

Dresden Database Systems Group

PROGNOSEN

- Schätzung zukünftiger und vergangener Werte
- Genauer durch Kenntnis externer Einflüsse

[Bildquelle: ls1.wettercomassets.com/img/cms/chameleon/mediapool/thumbs/0/c8/article_content_inline_full_1333514399_Regenradar_1503.jpg]

MERKMALSAUSWAHL

- Verbesserte Interpretierbarkeit der Ergebnisse
- Modellbildung beschleunigen
- Präzision der Vorhersage verbessern
- Geringerer Datenverarbeitungsaufwand
- Bei Wechselwirkungen: filtern weniger geeigneter Merkmale

Gliederung

Dresden Database Systems Group

1) **Methoden der Merkmalsauswahl**

2) Implementierungen

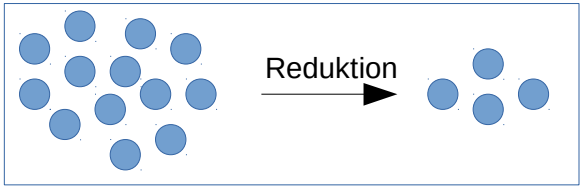
3) Evaluation

4) Zusammenfassung und Ausblick

Methoden der Rückwärtselimination

Dresden Database Systems Group

AUSWAHL DURCH REDUKTION



- Start mit allen Merkmalen
- Bewertungen bilden
- Merkmale mit bester Bewertung erhalten, Rest eliminieren

KORRELATIONSBASIERTE AUSWAHL

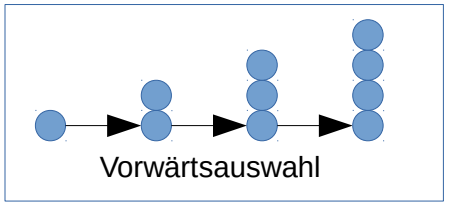
- Am wenigsten mit Zielgröße korrelierende Einflüsse entfernen
- Vorteil: Findet zuverlässig am höchsten korrelierten Einfluss
- Nachteile:
  - Wählt redundant aus, wenn gewählte Merkmale ähnlich gut korrelieren
  - Verwirft möglicherweise wichtige, gering korrelierte Einflüsse

Vorwärtsauswahl

Dresden Database Systems Group

FORWARD STEPWISE SELECTION

- Schrittweise am meisten mit Residuum korreliertes Merkmal wählen
- Mit bisheriger Auswahl Antwort modellieren, Residuum bilden
- Nachteil: Keine Gewichtung, binäre Auswahl
- Vorteil: Schnell



FORWARD STAGEWISE SELECTION

- Koeffizienten des am stärksten korrelierten Merkmals minimal erhöhen
- Mit aktueller Auswahl neues Residuum berechnen
- Nachteil: Langsam
- Vorteil: Gewichtet

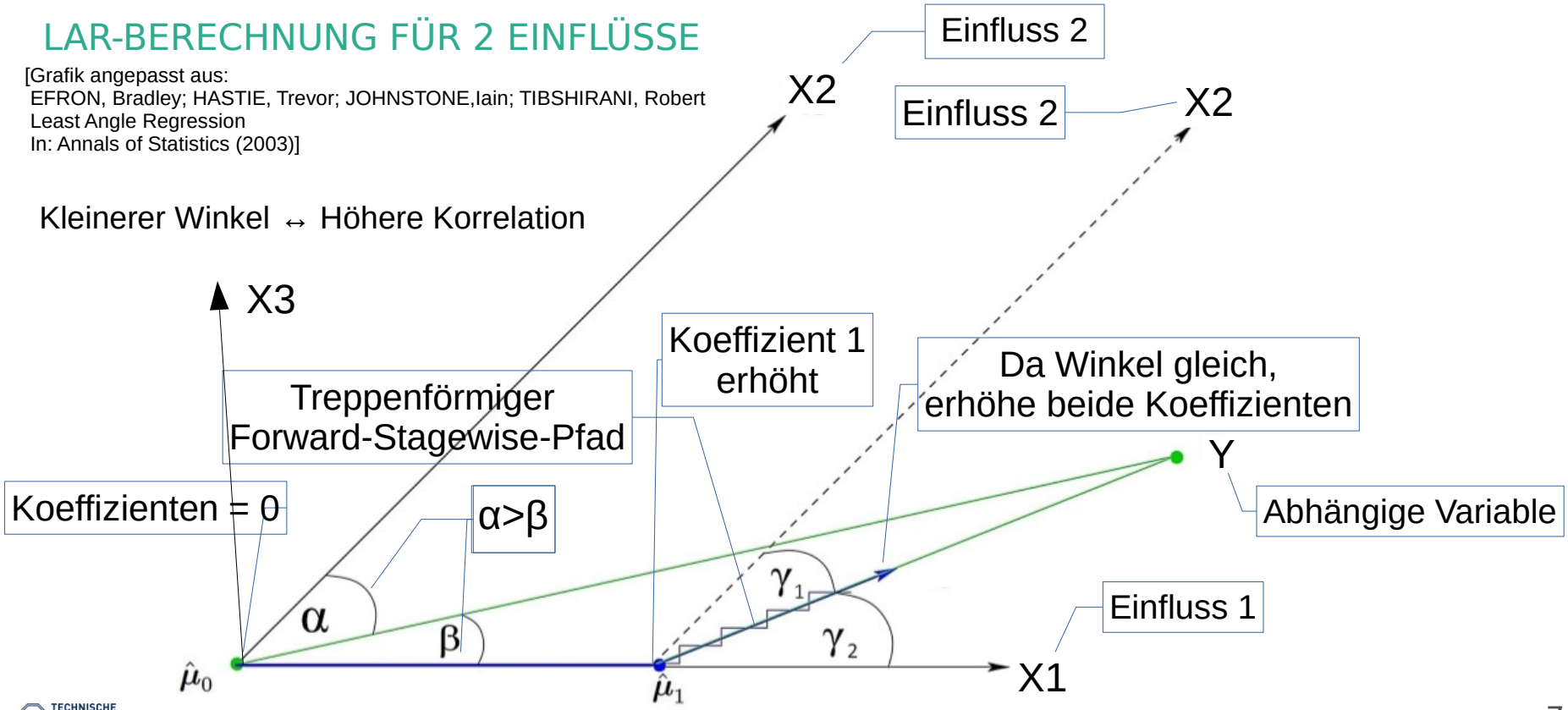
Least Angle Regression (LAR)

Dresden Database Systems Group

LAR-BERECHNUNG FÜR 2 EINFLÜSSE

[Grafik angepasst aus:
EFRON, Bradley; HASTIE, Trevor; JOHNSTONE,Iain; TIBSHIRANI, Robert
Least Angle Regression
In: Annals of Statistics (2003)]

Kleinerer Winkel ↔ Höhere Korrelation



Erweiterung von LAR zur Selektion

Dresden Database Systems Group

LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO)

- L1-Schranke (Begrenzung des Absolutbetrags der Summe der Koeffizienten)
- Suche Regressionsfunktion mit L1-Norm und minimaler Summe der Fehlerquadrate

LEAST ANGLE REGRESSION AND SELECTION (LARS)

- Berechnet LAR
- Wenn Koeffizient auf Null geschätzt, entfernen
- Wähle Zwischenbelegung unter Beachtung der L1-Schranke

- Vorteil: Gewichtete, reduzierte Merkmalsmenge performant berechenbar

Methoden der Vorwärtsauswahl

Dresden Database Systems Group

ZUSSAMENFASSENDER VERGLEICH DER DREI FORWARD-SELECTION-ANSÄTZE

Algorithmus	gewichtet	performant
Forward-Stepwise	Nein	Ja
Forward-Stagewise	Ja	Nein
LARS-LASSO	Ja	Ja

Gliederung

Dresden Database Systems Group

1) Methoden der Merkmalsauswahl

2) Implementierungen

3) Evaluation

4) Zusammenfassung und Ausblick

Tresholdbasierte Auswahl

Dresden Database Systems Group

ZIEL

Filterung zur Auswahl durch Treshold, Vergleich Korrelationsfilter mit LARS-LASSO

ABLAUF

- Berechne Korrelationskoeffizienten aller Merkmale
- Wähle Merkmale mit LARS-LASSO-Koeffizienten
- Erstelle Bewertung: Teilen jedes Koeffizienten durch den größten
- Filtere Merkmale mit Bewertung unter Treshold-Wert für LARS-LASSO und Korrelation

Iterative Auswahl

Dresden Database Systems Group

ZIEL

LARS-LASSO als einzige Methode, durch Mehrfachanwendung härtere Auswahl

ABLAUF

- Wähle Merkmale mit LARS-LASSO-Koeffizienten
- Wende LARS-LASSO auf gewählte Merkmale erneut an
- Bis Merkmalsmenge gleich bleibt → finale Merkmalsmenge gefunden

- Ausführen bis n Merkmale übrig, Beschränkung der Iterationen

Iterativ validierte Auswahl

Dresden Database Systems Group

ZIEL

Minderung des Einflusses von Multikorrelationen, Nachvollziehbarkeit der Verbesserung

ABLAUF

- Wähle Merkmal mit höchsten Koeffizienten nach LARS-LASSO-Auswahl
- Berechne Vorhersage mit Modell bisher gewählter Einflüsse
- Wenn nRMSE reduziert → Merkmal in finale Merkmalsmenge aufnehmen
- Wende LARS-LASSO erneut an, ohne bereits gewählte Merkmale zu beachten
- Voraussetzungen für Effizienz:
  - schnelles Modellbildungsverfahren zur Validierung
  - Obergrenze zu wählender Merkmale
  - Alternativ: zuerst LARS-LASSO zur Auswahl der zu betrachtenden Merkmale

Auswahl mit Clustering

Dresden Database Systems Group

ZIEL

Aus jeder korrelierten Merkmalsmenge einzeln die besten Einflüsse extrahieren

ABLAUF

- Bilde hierarchisches Cluster
- Nimm alle Merkmale in Cluster der Abhängigen Variablen in finale Ergebnismenge
- Wende LARS-LASSO mit bisheriger Ergebnismenge und je einem Cluster an
- Ergänze Auswahl durch von LARS-LASSO gewählte Merkmale
- Nachteil: durch Clusterbildung sehr langsam

Gliederung

1) Methoden der Merkmalsauswahl

2) Implementierungen

3) Evaluation

4) Zusammenfassung und Ausblick

Setup

Dresden Database Systems Group

DATEN

- DREWAG: 28 natürliche, 2 berechnete Merkmale, 19 (+2) Lastgänge
- GEFCOM: 12 natürliche, 4 berechnete Merkmale, 3 Lastgänge

GBM: Gradient Boosting Machines
MARS: Multiple Adaptive Regression Splines
MLR: Multiple Linear Regression

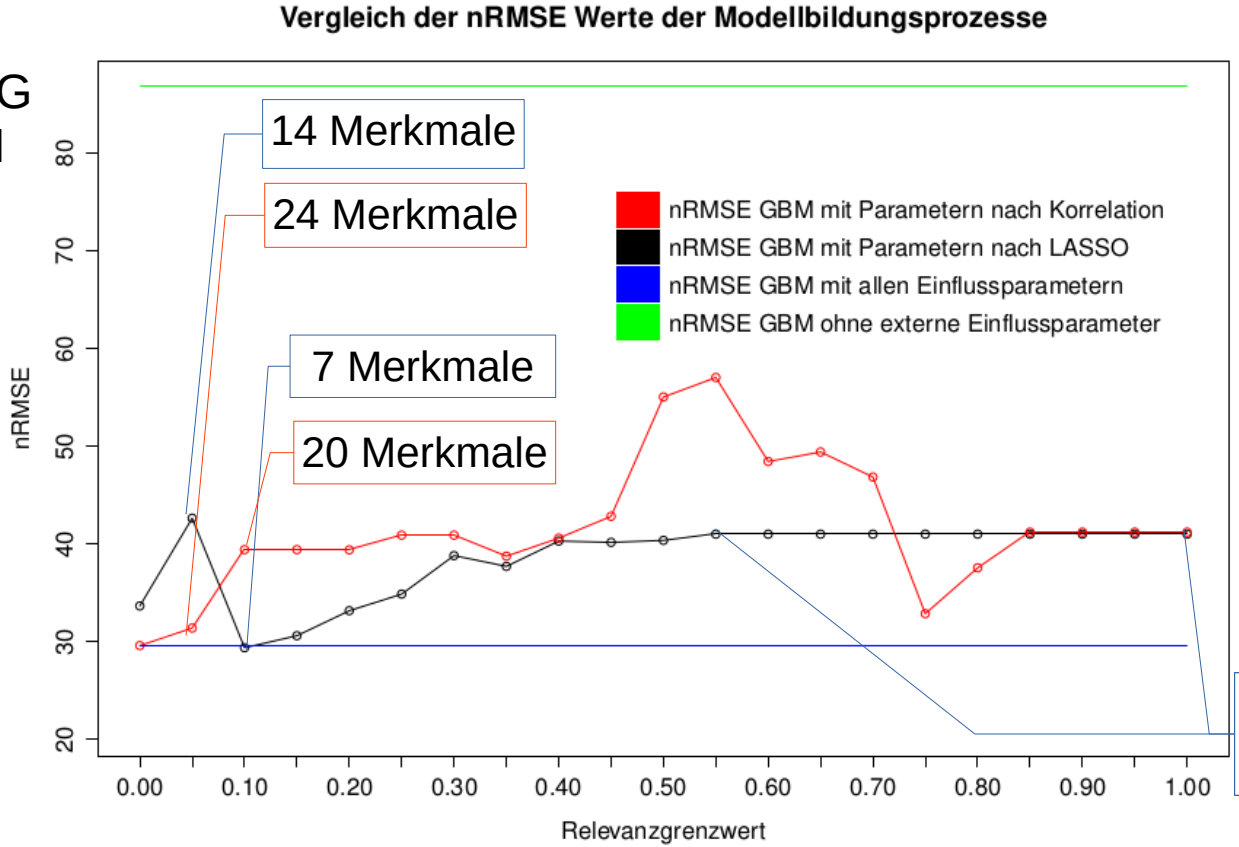
TESTBEDINGUNGEN

- Primäres Fehlermaß ist der **n**ormierte **R**oot-**M**ean-**S**quare-**E**rror
- Zeitpunkte, an denen der Lastgang den Wert Null hat, werden ignoriert
- Es werden mindestens MARS, MLR und GBM-Modelle gebildet
- Verglichen wird anhand von Modellbildungsdauer und Vorhersagegenauigkeit
- Angefügte Zufallsspalten verstärken Wirkung

Tresholdbasierte Auswahl I

Dresden Database Systems Group

DREWAG mit GBM



Die 14 Merkmale sind:

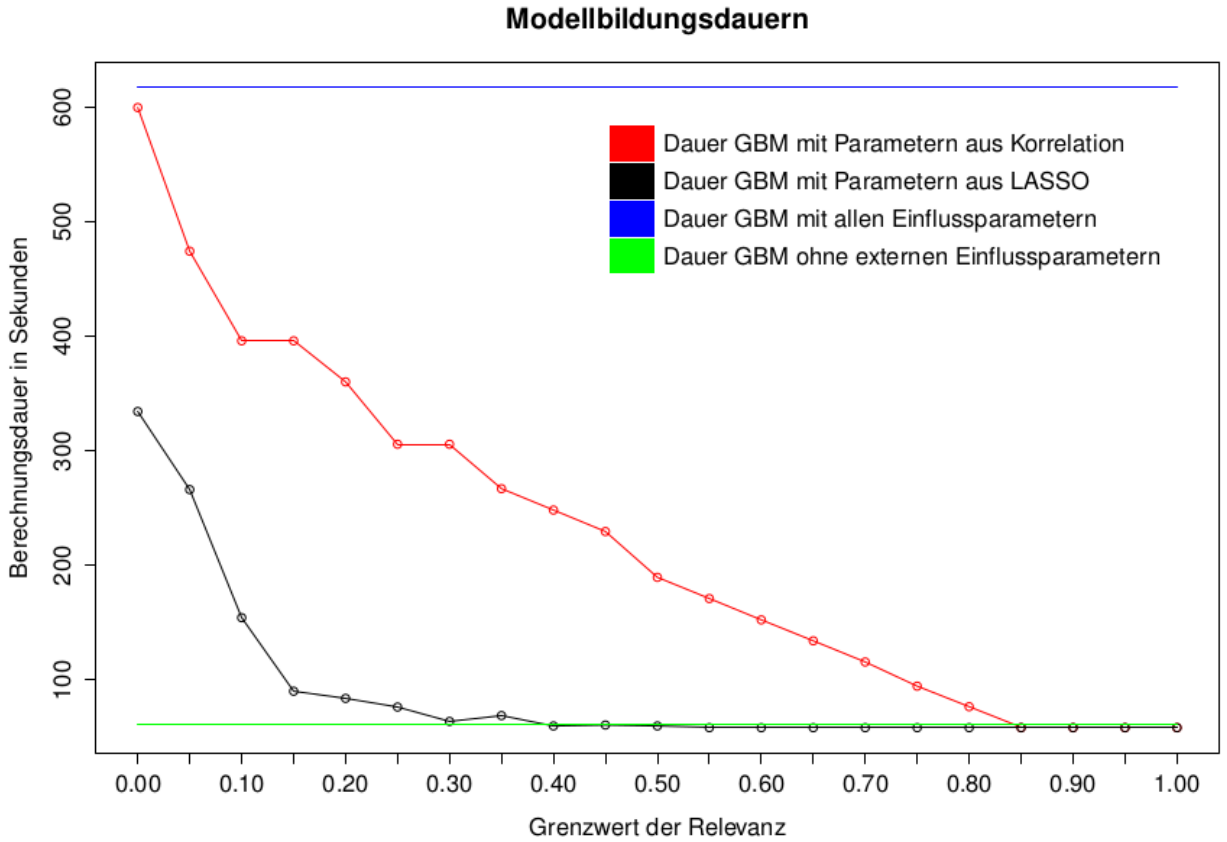
- Erdboden-Temperatur (20cm)
- Niederschlagshöhe
- Kurzwellige Strahlung
- Luftdruck
- Luftfeuchtigkeit
- Lufttemperatur (2)
- Relative Feuchte (2)
- Sonnenscheindauer

Globale Kurzwellen-strahlung

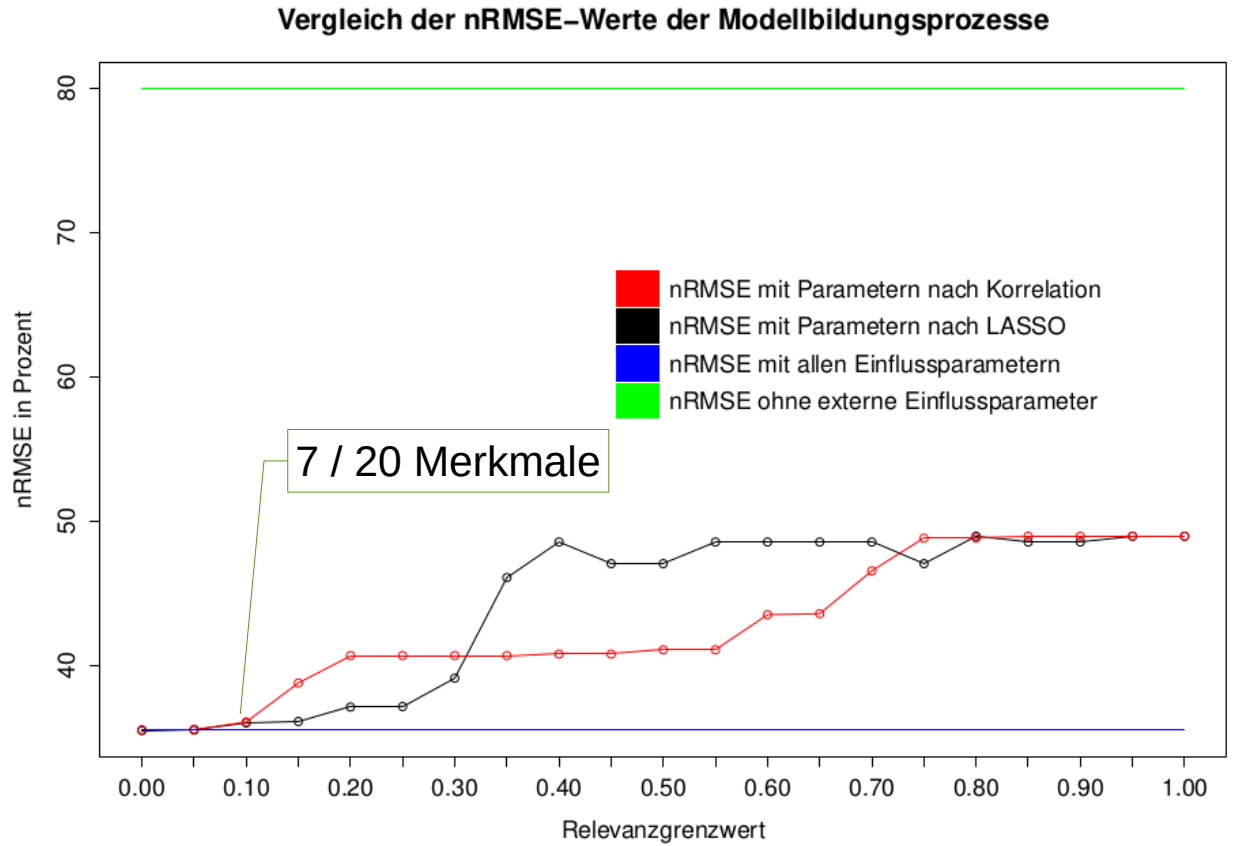
Tresholdbasierte Auswahl II

Dauer DREWAG mit GBM ohne zusätzliche Spalten →

mit 150 zusätzlichen Zufallsspalten: Mit LASSO 1 Minute statt >1h (durchschnittlich)



Tresholdbasierte Auswahl III



GEFCOM
mit MLR

Iterative Auswahl (GBM)

Dresden Database Systems Group

GEFCOM MIT 150 ZUFALLSSPALTEN

- Erste Ausführung wählt 9 der 16+150 Merkmale, Merkmalsmenge bleibt erhalten
- NRMSE von 31 auf 35 gestiegen (ohne externe Einflüsse: 80)
- Dauer von 4300s auf 230s gesunken (ohne externe Einflüsse: 70s)

DREWAG OHNE 150 ZUFALLSSPALTEN

Iterationen	0	1	n
Merkmale	30	17	11
Dauer	550	300	200
nRMSE	30	34	44

Zum Vergleich:
Ohne externe Einflüsse:
Dauer: 60s
nRMSE: 82

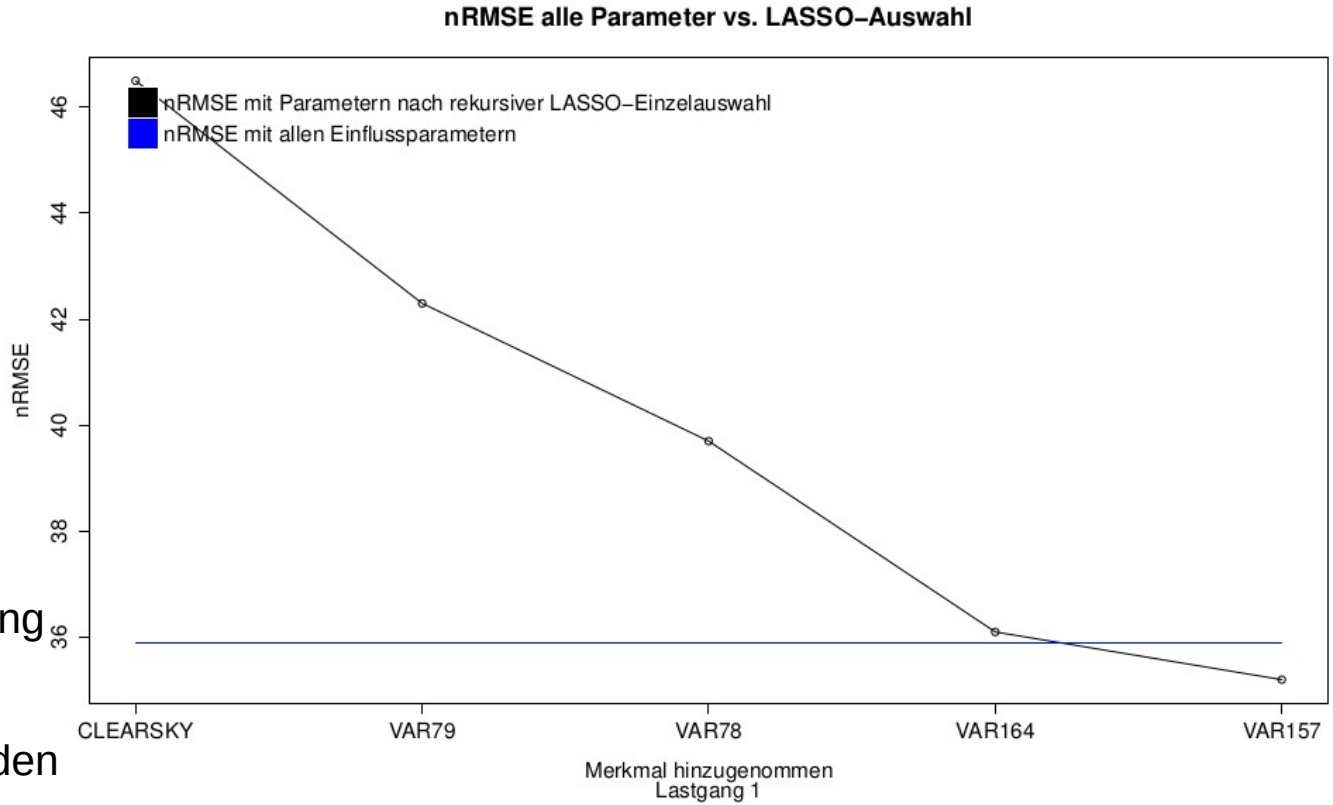
Iterativ validierte Auswahl (MLR)

Dresden Database Systems Group

GEFCOM
Lastgang 1
mit MLR und
150 Zufallsspalten

Nur für:
Schnelle Modell-bildungsverfahren
→ Dauer bis finales Modell 15-20mal so lang wie ohne Auswahl

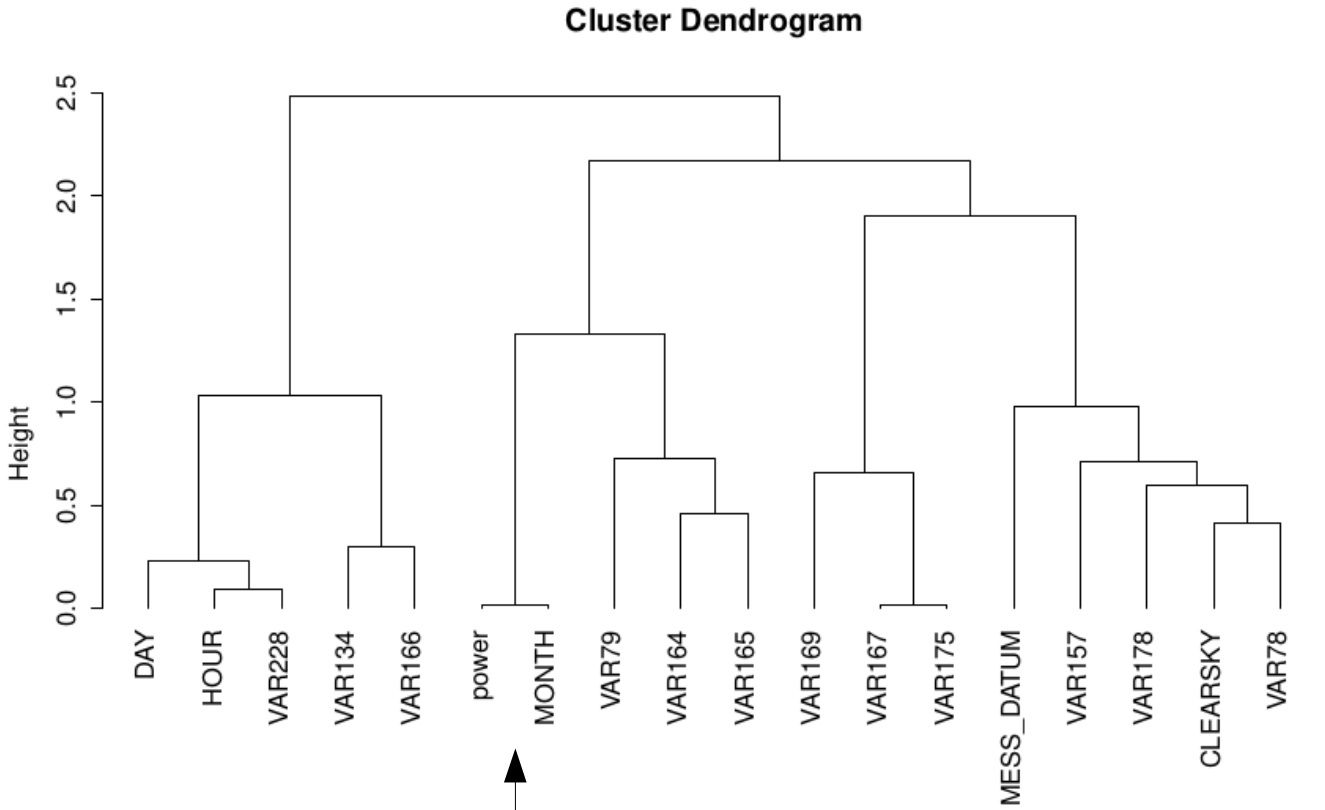
Mit Auswahl: 7 Sekunden



Auswahl mit Clustering

Dresden Database Systems Group

GEFCOM-Dendrogram



Durch Clustern:
- zeitintensiv
- fehleranfällig

Gliederung

1) Methoden der Merkmalsauswahl

2) Implementierungen

3) Evaluation

4) Zusammenfassung und Ausblick

Zusammenfassung und Ausblick

Dresden Database Systems Group

ZUSAMMENFASSUNG

- LARS-LASSO vielseitig nutzbar
- Einfachste Reduktion durch iterative Vorwärtsselektion
- Beste Resultate mit schnellen Verfahren und Verifizierung der Auswahl je Schritt
- Findung eines Tresholds kann Auswahl verschärfen und optimieren

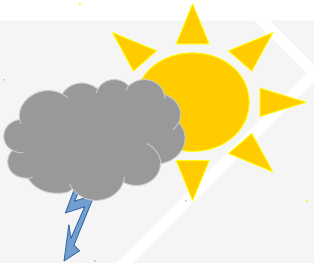
AUSBLICK

- Erweiterungen zum systematischen Probieren
- Beispiel: LARS-LASSO 2x anwenden, abgewählte Merkmale prüfen

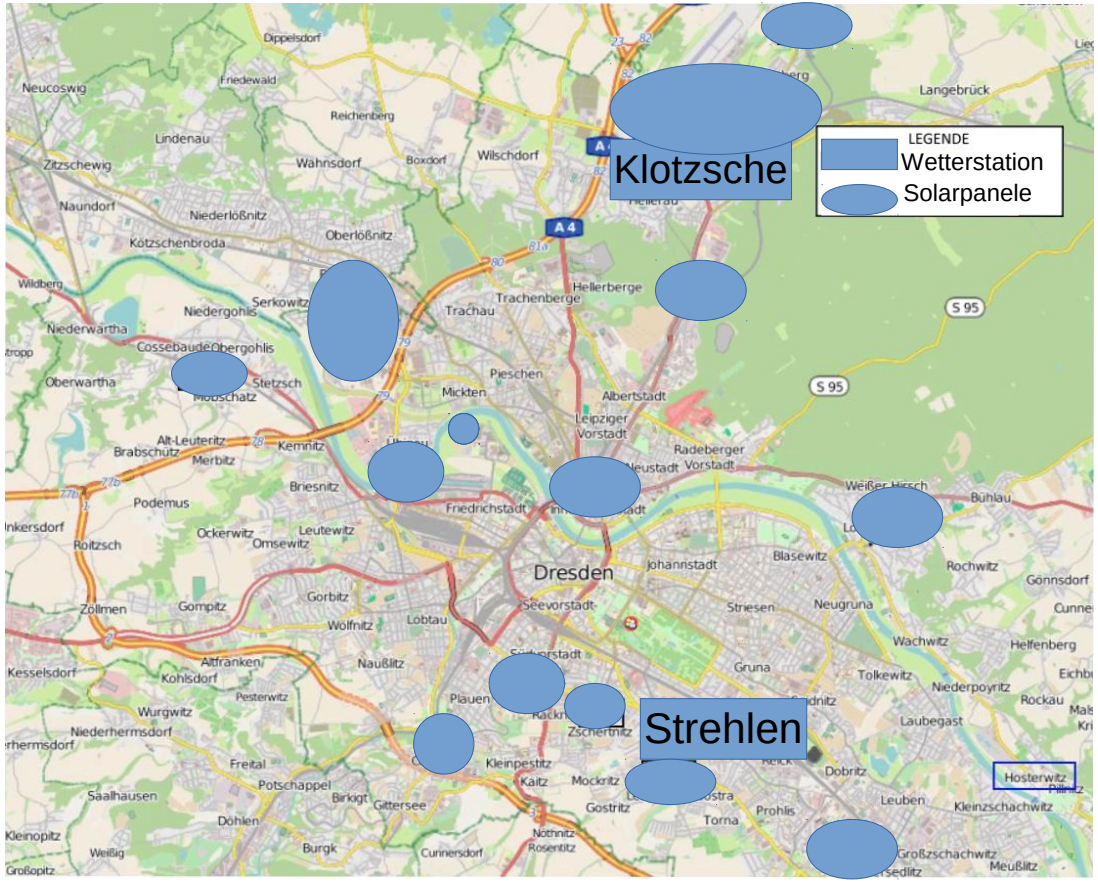
- Kombination mit korrelationsunabhängigen Methoden

- Elasticnet nutzen (L1+L2-Schranke) → zusätzliche Gewichtung

Besten Dank für Ihre Aufmerksamkeit



Lage DREWAG Wetter & Lastgang



Karte bereitgestellt durch [www.openstreetmap.de/karte.html]

Motivation II

Dresden Database Systems Group

WARUM NICHT EINFACH ALLE MERKMALE NUTZEN ?

- Wichtung der Einflüsse schwierig
- Modellbildung langsam
- Interpretation erschwert
- Hoher Datenverarbeitungsaufwand

→ Auswahl wichtiger Einflüsse notwendig