Cleaning von Zeitreihen

Von der Anomalienerkennung zur Anomalienreparatur

Jose Rodriguez Parra Flores
Klaus-Johan Ziegert

16. September 2019

Gliederung

Einführung

Motivation

Messgeräte liefern unzuverlässige Daten

- GPS Tracker sind nahe von Gebäuden unzuverlässig
- Sensoren sind empfindlich gegenüber äußere Einflüsse
  - Z.B. starker Fall der Temperaturen bei einem Windzug

Abbildung: GPS-Tracking auf dem Campus der Tsinghua Universität [1]

Motivation

Umgang von unzuverlässigen Daten mit Anomalienerkennung

1. Unzuverlässige Datenpunkte entfernen
   - Ausreißer werden entfernt 🙂
   - Entfernen aufeinanderfolgende Fehler machen Ergebnis unbrauchbar **oder** werden als solche ggf. nicht entfernt ☹
2. Unzuverlässige Datenpunkte reparieren
   - Einzelne Ausreißer werden leicht korrigiert 😐
   - Aufeinanderfolgende Fehler werden zu stark verändert (In der Praxis liegen die Messungen nahe bei den korrekten Werten) 😐

Motivation

Hinzunahme von korrekt markierten Werten

1. Markierung durch den Benutzer
   - Z.B. markiert der Benutzer in beliebigen Zeitabständen seinen aktuellen Standort
2. Präzise Messgeräte liefern in längeren Zeitabstände korrekte Werte

Zielsetzung

Ziel der Arbeit

1. Berücksichtigung der markierten Werte in der Anomalienerkennung
   - Aufeinanderfolgende Fehler sollen besser abgeschätzt werden
2. Anomalienreparatur mit den Minimum-Change-Prinzip vereinbaren
   - Keine drastische Veränderungen der Messwerte
3. Neue Anomalienreparatur hinsichtlich Berechnungslaufzeit, Ergebnisgenauigkeit usw. optimieren
4. Neue Anomalienreparatur mit unterschiedlichen Einstellungen mit den anderen Verfahren empirisch vergleichen

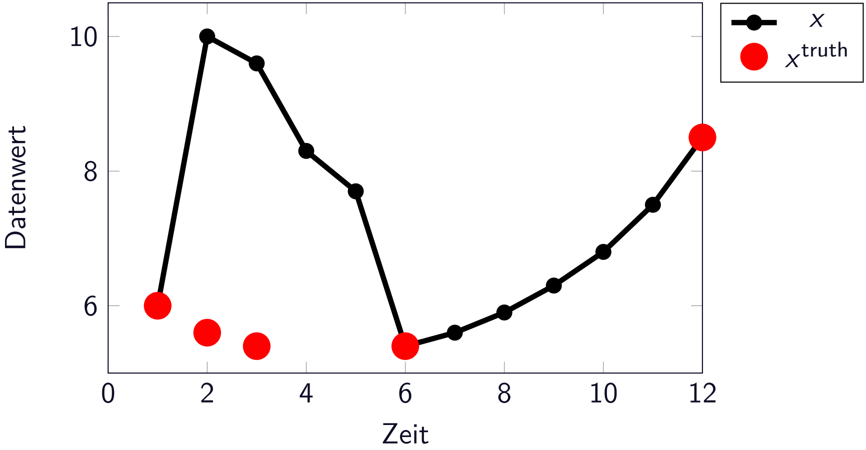
Grundlagen

Problemstellung

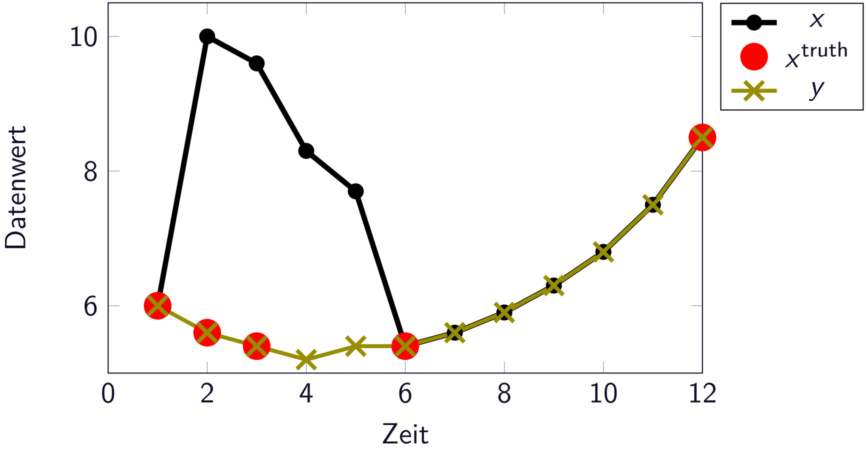
Zeitreihenreparatur

- Gegeben:
  - Unzuverlässige Messung $x = x[1], \ldots, x[n]$
  - Unvollständige, aber dafür ausschließlich korrekte Messung x^{truth}
  - (Nur bei der Evaluierung: vollständige, korrekte Messung $x^{\text{truth*}}$)
- Gesucht:
  - Reparatur y mit minimalen RMS-Fehler $\Delta(x^{\text{truth*}}, y)$
  - $\Delta(x^{\text{truth*}}, y) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(x_i^{\text{truth*}} - y_i\right)^2}$

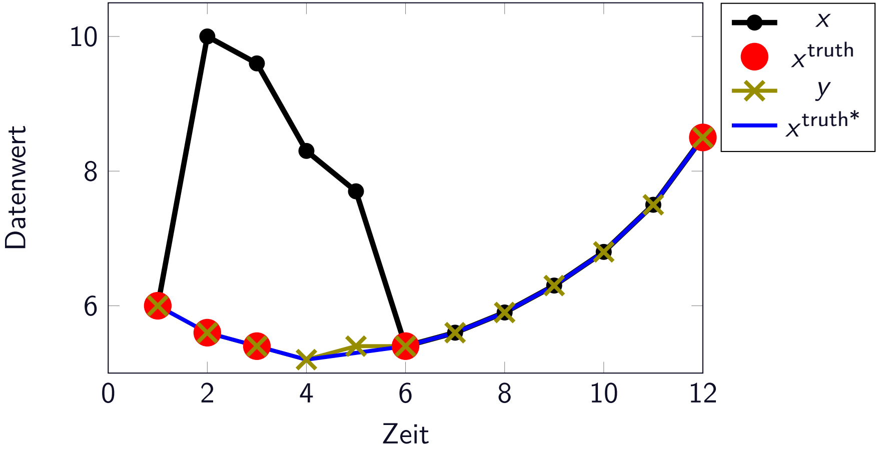
Anomalien: Beispiel Eingabe



Anomalien: Beispiel Eingabe & Reperatur



Anomalien: Beispiel Eingabe & Reparatur



Anomalien: Beispiel Zahlen & Bewertung

Zeitreihen vom Beispiel

- $x = \{6, 10, 9.6, 8.3, 7.7, 5.4, 5.6, 5.9, 6.3, 6.8, 7.5, 8.5\}$
- $x^{\text{truth}} = \{6, 5.6, 5.4, \_\_, \_\_, 5.4, \_\_, \_\_, \_\_, \_\_, \_\_, 8.5\}$
- $y = \{6, 5.6, 5.4, \underline{5.2, 5.4}, 5.4, \underline{5.6, 5.9, 6.3, 6.8, 7.5}, 8.5\}$
- $x^{\text{truth*}} = \{6, 5.6, 5.4, \underline{5.2, 5.3}, 5.4, \underline{5.6, 5.9, 6.3, 6.8, 7.5}, 8.5\}$

Bewertung des Beispiels

$$\Delta(x^{\text{truth*}}, y) =$$

$$\sqrt{\frac{1}{12}((6-6)^2 + \cdots + (5.3-5.4)^2 + \cdots + (8.5-8.5)^2)} \approx 0.03$$

Anomalienreparaturen

Anomalien

- Wikipedia: Abweichung von der Regel
- Werte x_i mit Abweichung τ (Bsp. $\tau = 2\sigma$):
  $|x_i - x_i^{\text{truth}}| > \tau$

Reparatur durch Anomalienerkennung

Autoregressive Modell $AR(p)$

- Lineare Regression der letze p Werte:

$$x'_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t$$

- Reparatur:

$$y_t = \begin{cases} x'_t & \text{falls } \textit{kein Label und } |x'_t - x_t| > \tau \\ x_t & \text{sonst} \end{cases}$$

Reparatur durch Anomalienerkennung

Autoregressives exogenes Modell $ARX(p)$

- Exogenes Variabel y

$$y'_t = x_t + \sum_{i=1}^{p} \phi_i (y_{t-i} - x_{t-i}) + \epsilon_t$$

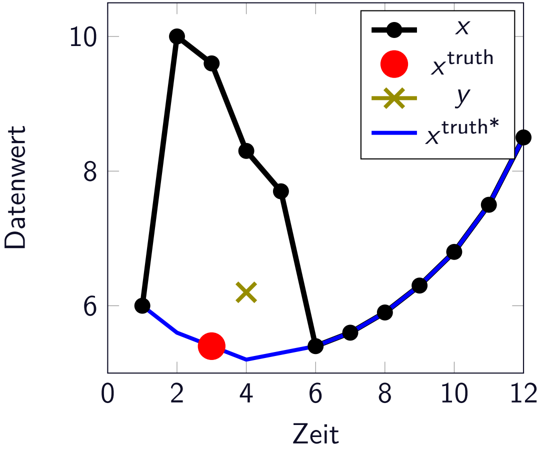
- y'_t Mögliche Reparatur:

$$y_t = \begin{cases} y'_t & \text{falls } \textit{kein Label und } |y'_t - x_t| > \tau \\ y_t & \text{sonst} \end{cases}$$

-

ARX Reparatur Beispiel

ARX(1) mit
$p = 1$, $\phi = 0.5$ und $\tau = 0.1$
$y'_4 = 8.3 + 0.5 \cdot (5.4 - 9.6)$
$y'_4 = 6.2$
$|6.2 - 8.3| = 2.1 > 0.1$
$y_4 = y'_4 = 6.2$



Andere Reparaturen

Gleitender Mittelwert

- $$y_j = \frac{1}{k} \sum_{i=1}^{k} x_{j-i} \quad j \in \{k, .., n\}$$

Andere Reparaturen

Gleitender Mittelwert

- $$y_j = \frac{1}{k} \sum_{i=1}^{k} x_{j-i} \quad j \in \{k, .., n\}$$

Exponentiell Gewichteten Gleitender Mittelwert (EWMA)

- $$v_j = \sum_{i=0}^{n} (\beta - 1)\beta^i x_{j-i} \quad j \in \{k, .., n\}$$

  as

  $$v_j = \beta v_{j-1} + (1 - \beta) x_j \quad mit \;\; V_0 = 0$$

Andere Reparaturen

Gleitender Mittelwert

- $$y_j = \frac{1}{k}\sum_{i=1}^{k} x_{j-i} \quad j \in \{k, .., n\}$$

Exponentiell Gewichteten Gleitender Mittelwert (EWMA)

- $$v_j = \sum_{i=0}^{n} (\beta - 1)\beta^i x_{j-i} \quad j \in \{k, .., n\}$$

  as

  $$v_j = \beta v_{j-1} + (1-\beta)x_j \quad mit \;\; V_0 = 0$$

SCREEN

-

Iterative Minimum Repairing

Intuitiver Ansatz von IMR

- ARX nutzt markierte Werte effizient, **aber** verändert die Werte zu drastisch.
- IMR Ansatz:
  1. Wende ARX an
  2. Wähle Reparaturwert aus mit minimalen Abstand zur Messung
  3. Wiederhole Prozedur
- Motivation: Reparierte Werte verbessern zukünftige Reparaturen

IMR = ARX + Minimum-Change-Prinzip

Eingabe: Messung x, markierte Werte x^{truth}, Ordnung p, Schwellenwert τ und max-num-iterations

Ausgabe: Reparatur y

$y^{(0)} \leftarrow \text{Initialize}(x, x^{\text{truth}})$

for $k \leftarrow 0$ **to** max-num-iterations **do**

  $\phi^{(k)} \leftarrow \text{Estimate}(x, y^{(k)})$

  $\hat{y} \leftarrow \text{Candidate}(x, y^{(k)}, \phi^{(k)})$

  $y^{(k+1)} \leftarrow \text{Evaluate}(x, y^{(k)}, \hat{y})$

  if $\text{Converge}(y^{(k)}, y^{(k+1)})$ **then**

    break

  end if

end for

return $y^{(k)}$

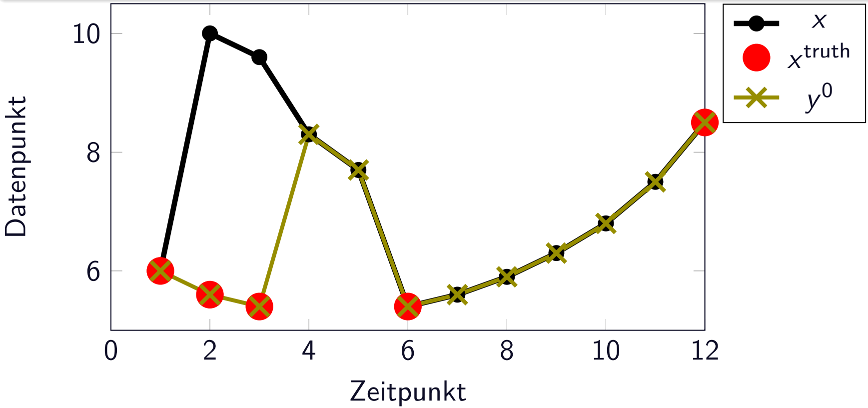
IMR: Initialisierung

Eingabe: Messung x, markierte Werte x^{truth}, Ordnung p, Schwellenwert τ und max-num-iterations
Ausgabe: Reparatur y
$y^{(0)} \leftarrow \text{Initialize}(x, x^{\text{truth}})$
for $k \leftarrow 0$ **to** max-num-iterations **do**
    $\phi^{(k)} \leftarrow \text{Estimate}(x, y^{(k)})$
    $\hat{y} \leftarrow \text{Candidate}(x, y^{(k)}, \phi^{(k)})$
    $y^{(k+1)} \leftarrow \text{Evaluate}(x, y^{(k)}, \hat{y})$
    if $\text{Converge}(y^{(k)}, y^{(k+1)})$ **then**
        break
    end if
end for
return $y^{(k)}$

IMR: Initialisierung

Initiale Reparatur

Initiale Reparatur $y^{(0)}$ ist Messung x und übernimmt die markierten Werte aus x^{truth}



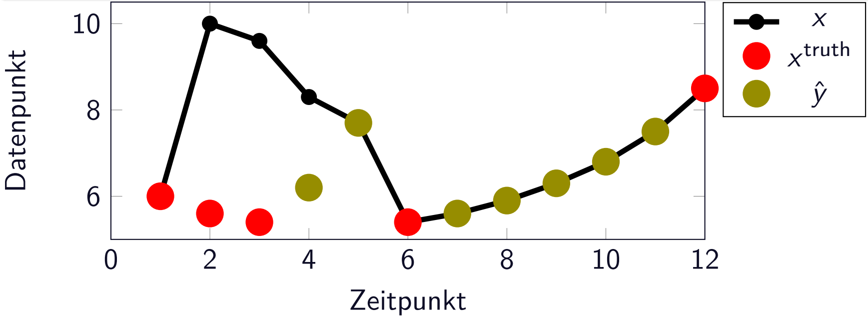
IMR: ARX auf aktuelle Reparatur anwenden

Eingabe: Messung x, markierte Werte x^{truth}, Ordnung p, Schwellenwert τ und max-num-iterations
Ausgabe: Reparatur y
$y^{(0)} \leftarrow \text{Initialize}(x, x^{\text{truth}})$
for $k \leftarrow 0$ **to** max-num-iterations **do**
    $\phi^{(k)} \leftarrow \text{Estimate}(x, y^{(k)})$
    $\hat{y} \leftarrow \text{Candidate}(x, y^{(k)}, \phi^{(k)})$
    $y^{(k+1)} \leftarrow \text{Evaluate}(x, y^{(k)}, \hat{y})$
    if $\text{Converge}(y^{(k)}, y^{(k+1)})$ **then**
        break
    end if
end for
return $y^{(k)}$

IMR: ARX auf aktuelle Reparatur anwenden

Kandidaten

- Parameterschätzung ϕ: aktuelle Reparatur $y^{(k)}$ wird als x^{truth} interpretiert.
- Reparatur sind Kandidaten $\hat{y}$ (angewendetes ARX kennt vorangegangene Aktualisierung nicht)



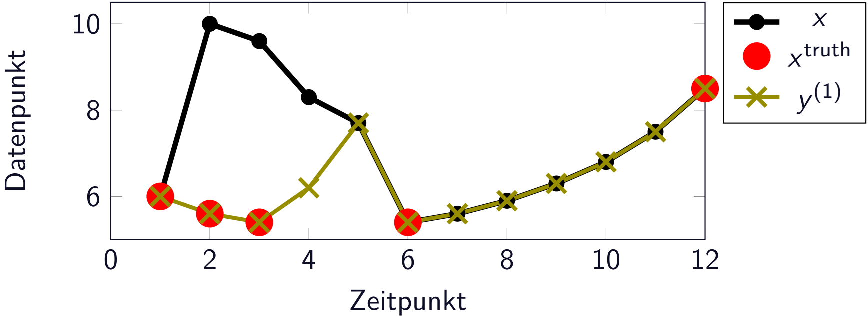
IMR: Minimum-Change

Eingabe: Messung x, markierte Werte x^{truth}, Ordnung p, Schwellenwert τ und max-num-iterations
Ausgabe: Reparatur y
$y^{(0)} \leftarrow \text{Initialize}(x, x^{\text{truth}})$
for $k \leftarrow 0$ **to** max-num-iterations **do**
    $\phi^{(k)} \leftarrow \text{Estimate}(x, y^{(k)})$
    $\hat{y} \leftarrow \text{Candidate}(x, y^{(k)}, \phi^{(k)})$
    $y^{(k+1)} \leftarrow \text{Evaluate}(x, y^{(k)}, \hat{y})$
    if $\text{Converge}(y^{(k)}, y^{(k+1)})$ **then**
        break
    end if
end for
return $y^{(k)}$

IMR: Minimum-Change

Minimum-Change

- Zu geringe Änderungen werden herausgefiltert $|y_i^{(k)} - \hat{y}_i| > \tau$
- Geringste Änderung zu Messung x wird als Kandidaten herangezogen



IMR: Terminierung

Eingabe: Messung x, markierte Werte x^{truth}, Ordnung p, Schwellenwert τ und max-num-iterations

Ausgabe: Reparatur y

$y^{(0)} \leftarrow \text{Initialize}(x, x^{\text{truth}})$

for $k \leftarrow 0$ **to** max-num-iterations **do**

  $\phi^{(k)} \leftarrow \text{Estimate}(x, y^{(k)})$

  $\hat{y} \leftarrow \text{Candidate}(x, y^{(k)}, \phi^{(k)})$

  $y^{(k+1)} \leftarrow \text{Evaluate}(x, y^{(k)}, \hat{y})$

  if $\text{Converge}(y^{(k)}, y^{(k+1)})$ **then**

    break

  end if

end for

return $y^{(k)}$

IMR: Terminierung

Terminierung

- Zwei Möglichkeiten der Terminierung:
  - Maximale Anzahl der Iterationen wird erreicht
  - Konvergenz: Neue Reparatur $y^{(k+1)}$ ist gleich aktuelle Reparatur $y^{(k)}$
- Allgemeine Konvergenzfrage ist noch offen

Motivation von Matrix-Pruning IMR

Laufzeitproblem

- Parameterschätzung beansprucht viel Zeit
- Matrizen V und Z bestehen aus $y_i^{(k)} - x_i$:
  - wenige markierte Werte vorhanden
  - markierte Werte häufig identisch zur Messung
  - Reparaturwerte ändern sich nicht signifkant
  - $\rightarrow$ dünnbesetzte Matrizen
- Matrix-Pruning: Löschen von Zeilen mit 0en

Matrix Pruning IMR Beispiel

Beispiel

Beispiel

Evaluierung

Ordnung

Blank

- Faktenöasdkfjöasdöasdkfjöasd
- Fakten
- Fakten

Schwellenwert

Blank

- Fakten
- Fakten
- Fakten

maximale Anzahl von Iterationen

Blank

- Fakten
- Fakten
- Fakten

Markierungsrate

Blank

- Fakten
- Fakten
- Fakten

Schluss

Zusammenfassung und Ausblick

Zusammenfassung

- Was wurde getan?

Zusammenfassung und Ausblick

Zusammenfassung

- Was wurde getan?

Ausblick

- Wie könnten zukünftige Arbeiten aussehen?

Literatur I

- Shaoxu song - tsinghua university.
- Aoqian Zhang, Shaoxu Song, Jianmin Wang, and Philip S Yu.
  Time series data cleaning: From anomaly detection to anomaly repairing.
  Proceedings of the VLDB Endowment, 10(10):1046–1057, 2017.