StatStream : Statistical Monitoring of Thousands of Data Streams in Real Time

Yunyue Zhu,Dennis Sasha

Vorgetragen von Matthias Altmann

Mehrfache Datenströme

- Beispiel Luft und Raumfahrttechnik:
  - Space Shuttle der NASA messen annähernd mit 20000 Sensoren
  - Ergebnisse werden an Johnson Space Center,Houston zur Auswertung geschickt

Mehrfache Datenströme

- Beispiel Finanzwesen:
  - Es gibt allein 50000 Wertpapiere in den USA
  - Und jede Sekunde bis zu 100000 Kursangaben und minimale Kursänderungen

Wie sind diese riesigen Datenmengen kontrollierbar?

Wie kann man diese Menge an Daten kontrollieren

- Updates durch Einbringung von neuen Elementen,d.h. seltene Korrektur älterer Daten
- Daten als Sequenzen nicht Mengen behandeln
- Daten als nie endende Datenströme betrachten
- One Pass Algorithmen sind erwünscht
- Da die Auswertung (meist) qualitativ ist,ist das Opfern von Genauigkeit für Geschwindigkeit akzeptabel

StatStream

- Versuch der Lösung dieses Problems durch einen Vorschlag StatStream genannt
- Darstellung an Beispielproblemen aus dem Finanzsektor , jedoch jederzeit verallgemeinerbar
- Unterscheidung von funktionellen und algorithmischen Lösungsansätzen :

Funktionelle Lösungsansätze StatStream

- Art und Weise , wie Datenstromstatistiken in StatStream berechnet werden :
  - konstant , fortlaufend mit einem Zeitfenster v
  - Per Diskreter Fourier Transformations-Annäherung
  - Welche einen geringen Fehler besitzen kann
  - Wiederbesuchen der verfallenen Datenströme nicht notwendig

Algorithmische Lösungsansätze

Unterscheidung von 3 Zeitdauern:

- Zeitpunkt (timepoint)
- Basisfenster (basic window) :
  aufeinanderfolgende Zeitpunkte über welche das System einen inkrementellen Auszug aufrecht erhält
- Gleitendes Fenster (Sliding Window) :
  aufeinanderfolgende Basisfenster über welche der Nutzer Statistiken will

Basisfenster bieten Vorteile in Bezug auf Verzögerung und Berechnung

Zeitlicher Abstand

Es werden 3 Arten von zeitlichen Abständen unterschieden:

- Grenzsteinfenster (landmark windows)
- Gleitende Fenster (sliding windows)
- gedämpftes Fenster Modell (damped window model)

Hier wird das gleitende Fenster-Modell genutzt

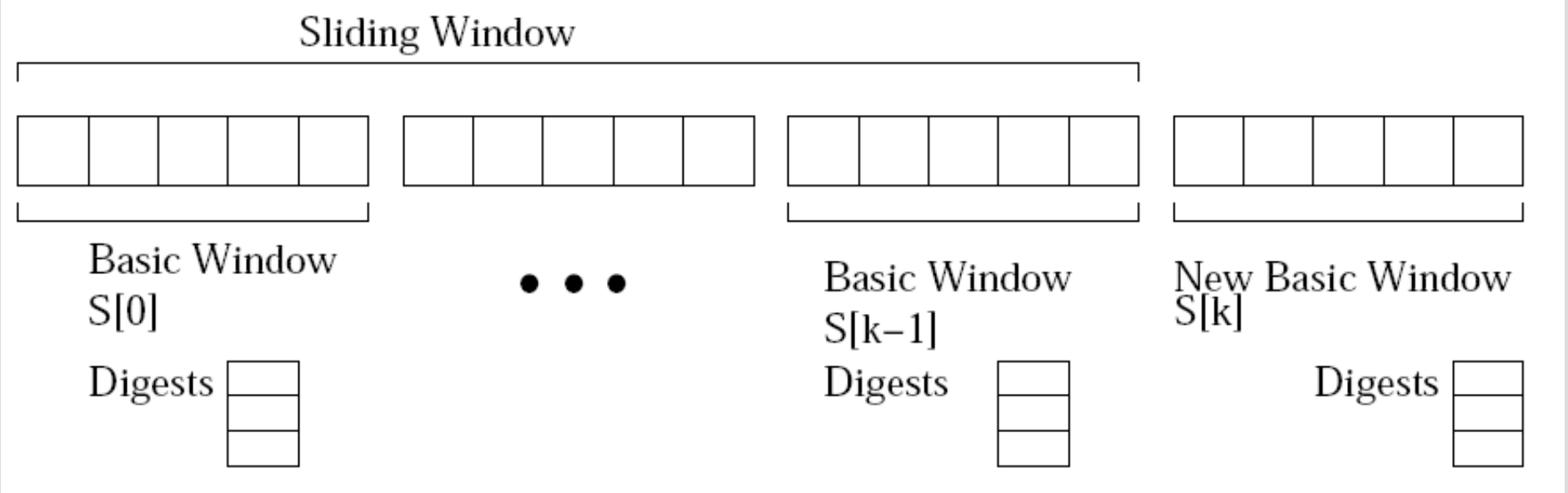
Statistiken zum Überwachen

Welche Statistiken wollen wir überwachen :

- Einzelne Datenstromstatistiken, wie
  - Durchschnitt
  - Standardabweichung
- Korrelationskoeffizienten
- Autokorrelationen der Serien
- Beta

Statistiken über gleitende Fenster

- 1 gleitendes Fenster wird in gleichmässig kleinere Fenster (Basisfenster) unterteilt ,
- Es werden Auszüge von Basisfenstern und Gleitenden Fenstern gehalten



Statistiken auf einzelnen Datenströmen

Hier nun ein Beispiel für eine Statistik auf einem einzelnen Datenstrom:

moving average

- Sei S[0],S[1],...,S[k-1] eine Folge von Basisfenstern
- S[k] wird das neue Basisfenster , S[0] ist das Auslaufende
- Dann muss für den moving average

$$\sum_{new}(s) = \sum_{old}(s) + \sum S[k] - \sum S[0]$$

aufrecht erhalten werden

Statistiken über mehrere Datenströme

Korrelationsstatistiken
zwischen mehreren Strömen

- Wo wichtig: z.B. Im Pairs Trading
  Einsatz in grossen,wichtigen Wallstreetfirmen
- Worst Case : Berechnung aller paarweiser Ähnlichkeiten : proportional zu Anzahl der Zeitpunkte in jedem Fenster * alle Paare von Datenströmen

Korrelationsstatistiken

- Der Worst Case soll hier durch 2 Faktoren verhindert werden
  - Annäherung der Korrelationen von Datenstrompaaren per Diskreter Fourier Transformation von Basisfenstern und
  - durch eine Netzdatenstruktur die die Annäherungsberechnung für die meisten Paare verhindert

Diskrete Fourier Transformation

Diskrete Fourier Transformation einer Zeitsequenz x= $x_0, x_1, ..., x_{w-1}$
ist eine Sequenz X= $X_0, X_1, ..., X_{w-1}$ = DFT(x)
von komplexen Zahlen gegeben durch

$$X_F = \frac{1}{\sqrt{w}} \sum_{i=0}^{w-1} x_i e^{-j2\pi Fi/w} \qquad F = 0, 1, ..., w-1$$

Wobei $j = \sqrt{-1}$
Inverse Transformation :

$$x_i = \frac{1}{\sqrt{w}} \sum_{F=0}^{w-1} X_F e^{j2\pi Fi/w} \qquad i = 0, 1, ..., w-1$$

Eigenschaften einer DFT

Die Berechnung der ersten paar DFT Koeffizienten haben den höchsten Aufwand :

diese Koeffizienten erfassen die Rohform der Zeitserie am Besten

Wenn wir diese Koeffizienten nutzen könnten , wären unsere Zeitserie gut angenähert und komprimiert dargestellt

Korrelationsstatistiken

- Korrelationen und der Beta können aus dem **Skalarprodukt** zweier Ströme berechnet werden
- Das Skalarprodukt über die gleitenden Fenster wird wieder auf die Basisfenster zurückgeführt
- Für Skalarprodukt wichtig : Basisfenster können synchronisiert oder zeitverzögert sein
- Hier soll nur die synchronisierte Variante betrachtet werden

Skalarprodukt mit angepassten Fenstern

Annäherung der xi und yi von 2 Basisfenstern mit Funktionsfamilien

$$x_i \approx \sum_{m=0}^{n-1} c_m^x f_m(i), y_i \approx \sum_{m=0}^{n-1} c_m^y f_m(i) \quad i = 1, \ldots, b$$

Berechnung des Skalarprodukts der angenäherten Basisfenster:

$$= \sum_{m=0}^{n-1} \sum_{p=0}^{n-1} c_m^x c_p^y \Big(\sum_{i=1}^{b} f_m(i) f_p(i) \Big)$$

Skalarprodukt mit angepassten Fenstern

$$= \sum_{m=0}^{n-1} \sum_{p=0}^{n-1} c_m^x c_p^y \left(\sum_{i=1}^{b} f_m(i) f_p(i) \right)$$

$$= \sum_{m=0}^{n-1} \sum_{p=0}^{n-1} c_m^x c_p^y$$

W(m,p) , wobei W(m,p) vorberechnet werden kann

Wenn die Funktionsfamilie orthogonal ist , ist

W(m,p) = $\begin{cases} 0 & m \neq p \\ V(m) \neq 0 & m = p \end{cases}$

$$\Rightarrow \sum_{i=1}^{b} x_i y_i \approx \sum_{m=0}^{n-1} c_m^x c_m^y V(m)$$

Skalarprodukt mit angepassten Fenstern

$$\sum_{i=1}^{b} x_i y_i \approx \sum_{m=0}^{n-1} c_m^x c_m^y V(m)$$

- O(kn) für das Skalarprodukt 2er angepasster Serien
- Datenkompression und Einfügen zusätzlicher Daten möglich

Bem. Bei Skalarprodukten mit unangepassten Basisfenster wird ähnlich vorgegangen

Skalarprodukt mit angepassten Fenstern

Wie kommt nun die DFT ins Spiel :
Die Cosinus / Sinusfunktionsfamilie hat die für die Annäherungsfunktionen richtigen Eigenschaften

Wir erhalten für ein Basisfenster

$$x_i \approx \frac{1}{\sqrt{b}} \sum_{F=0}^{n-1} X_F e^{j2\pi Fi/b} \qquad i = 1, 2, ..., b$$

Zur Erinnerung:

$$x_i \approx \sum_{m=0}^{n-1} c_m^x f_m(i), y_i \approx \sum_{m=0}^{n-1} c_m^y f_m(i)$$

Inkrementelle DFT Berechnung

$$x_i \approx \frac{1}{\sqrt{b}} \sum_{F=0}^{n-1} X_F e^{j2\pi Fi/b} \qquad i = 1, 2, ..., b$$

Man kann nun die DFT Koeffizienten (X_F oben) inkrementell berechnen

$$X_m^{new} = e^{\frac{j2\pi m}{w}} (X_m^{old} + \frac{x_w - x_0}{\sqrt{w}}) \qquad m = 1, ..., n$$

wobei x_0 ein auslaufendes x und x_w ein neues x ist

E/A Leistung

E/A-Kosten für gespeicherte Werte:

mit DFT-Annäherung

$$\frac{N_s kSizeof(float)(2+2n)}{PageSize}$$

fehlerfrei: $\frac{N_s kSizeof(float)b}{PageSize}$

Verbesserung : Verhältnis $\frac{b}{2+2n}$

Bem. die 2 in der Formel = die 2 Nicht-DFT-Elemente der Auszugsdaten : Die Summe und die Summe der Quadrate der Zeitserien in jedem Basisfenster

Überwachen von Korrelationen zwischen Datenströmen

- Hier nun ein Weg um schnell hohe Korrelationen zwischen Statistiken aufzudecken

Dafür benötigt man :
Normalisierung einer Serie über gleitende Fenster der Grösse w: $x_1, x_2, ..., x_w$ =

$$\hat{x}_i = \frac{x_i - \overline{x}}{\sigma_x}, \quad i = 1, 2, ..., w$$ wobei $$\sigma_x = \sqrt{\sum_{i=1}^{w} (x_i - \overline{x})^2}$$

Überwachen von Korrelationen zwischen Datenströmen

Man kann den Korrelationskoeffizienten nun auf die euklidische Distanz zurückführen

Euklidische Distanz zwischen 2 Zeitserien $x_1, \ldots, x_w$ und $y_1, \ldots, y_w$ =

$$d(x,y) = \sqrt{\sum_{i=1}^{w} (x_i - y_i)^2}$$

Überwachen von Korrelationen zwischen Datenströmen

Es gilt

$$corr(x,y) = 1 - \frac{1}{2}d^2(\hat{x},\hat{y})$$

$$corr(x,y) \geq 1 - \epsilon^2 \Rightarrow d_n(\hat{X},\hat{Y}) \leq \epsilon$$

$$corr(x,y) \leq -1 + \epsilon^2 \Rightarrow d_n(-\hat{X},\hat{Y}) \leq \epsilon$$

wobei $d_n(\hat{X},\hat{Y})$ die Euklidische Distanz zwischen den Zeitserien $\hat{X}_1, \hat{X}_2, ..., \hat{X}_n$ und $\hat{Y}_1, \hat{Y}_2, ..., \hat{Y}_n$ ist

Überwachen von Korrelationen zwischen Datenströmen

Man erhält eine Obermenge stark korrelierender Paare die nicht false negative sind

Schauen wir uns die normalisierten DFT Koeffizienten etwas näher an:

Überwachen von Korrelationen zwischen Datenströmen

Für alle normalisierten DFT Koeffizienten einer Sequenz $x_1, x_2, ..., x_w$ gilt

$$|\hat{X}_i| \leq \frac{\sqrt{2}}{2}, \quad i = 1, ..., n, n < w/2$$

=>

- der DFT Feature Raum ist ein Würfel mit Durchmesser $\sqrt{2}$
- Er hat 2n Dimensionen

Das wird genutzt :

Überwachen von Korrelationen zwischen Datenströmen

- Man unterteilt den Würfel in $(2\lceil\frac{\sqrt{2}}{2\epsilon}\rceil)^{\hat{n}}$ gleichgrosse,kleinere Würfel mit dem Durchmesser ϵ
- Zur Indexierung nutzt man die ersten $\hat{n},\ \hat{n} \leq 2n$ DFT- Koeffizienten
- So wird jeder Datenstrom auf eine Zelle basierend auf seine ersten $\hat{n}$Koeffizienten abgebildet

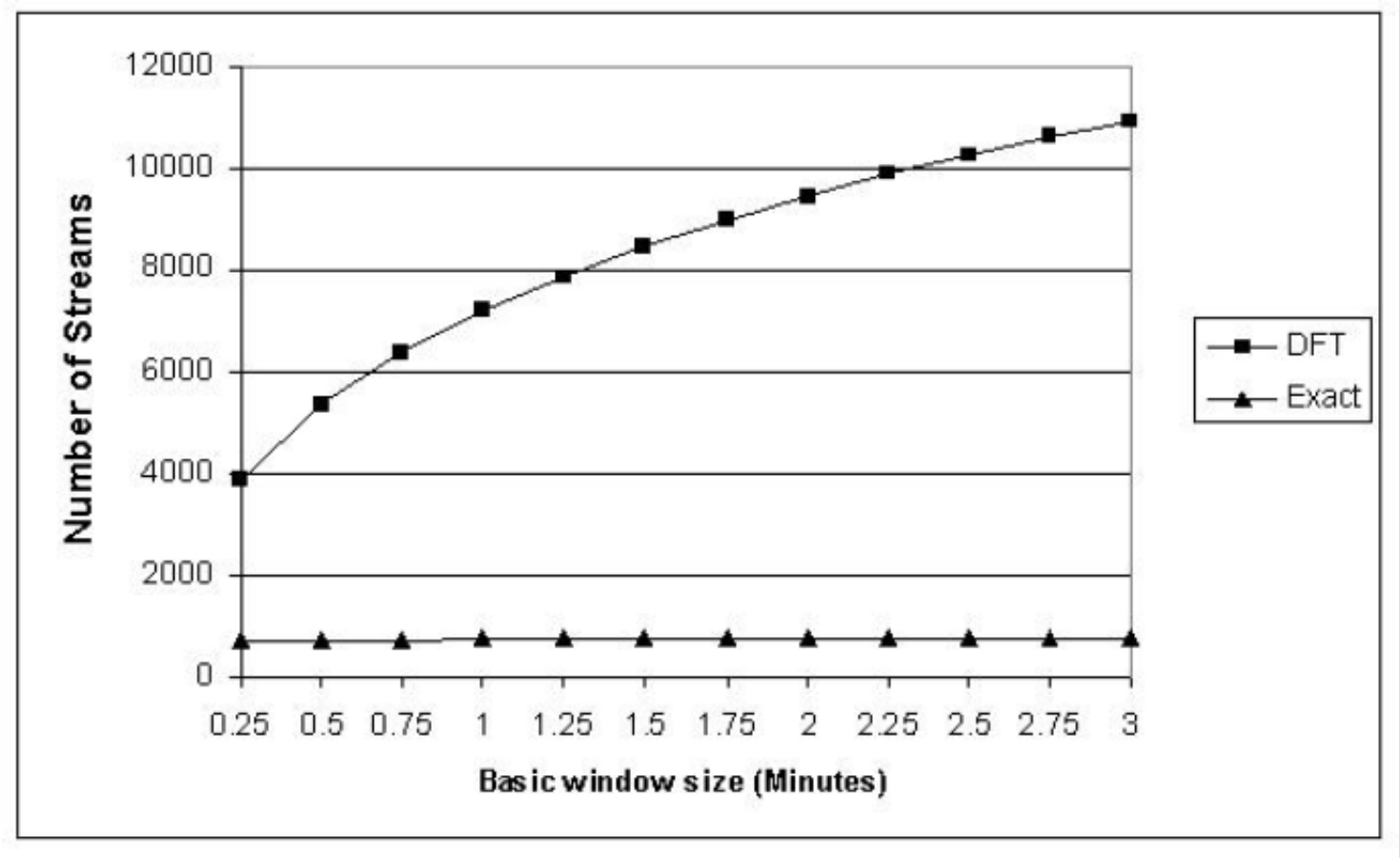
Überwachen von Korrelationen zwischen Datenströmen

Beispiel Strom x wird auf eine Zelle $(c_1, c_2, ..., c_{\hat{n}})$ gehasht :

- Ströme deren Korrelationskoeffizient > Grenzwert $1 - \epsilon^2$ liegen in direkter Nachbarschaft zu $(c_1, c_2, ..., c_{\hat{n}})$

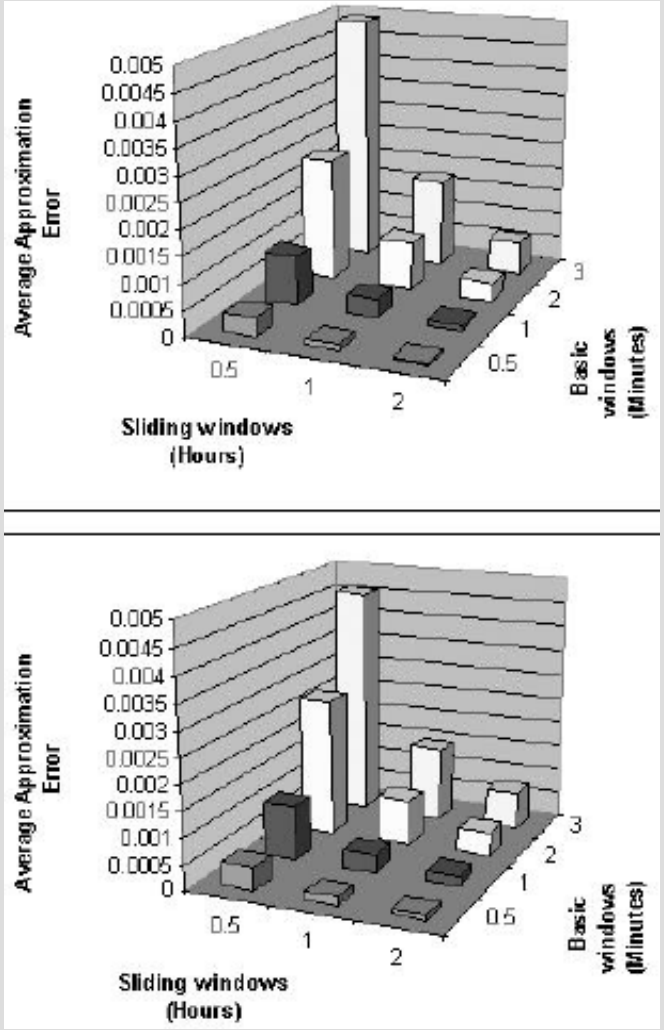
- Ströme deren Korrelationskoeffizient mit x < Grenzwert $-1 + \epsilon^2$ liegen auf Zellen benachbart zu $(-c_1, -c_2, ..., -c_{\hat{n}})$

Geschwindigkeitsmessung



Vergleich der Anzahl von Datenströmen die die DFT bzw. die fehlerfreie Methode halten kann

Precision Messung



Durchschnittlicher Annäherungsfehler für Korrelationskoeffizienten mit Basisfenster- / gleitende Fenstergrössen für synthetische (oben) und reelle (unten) Daten

Ähnliche Arbeiten

- Angenäherte Quantil-Berechnung für individuelle Datenströme
- Koevolutionäre Zeitserien
- Finden von Korrelationen zwischen Online – Sequenzen und einer indexierten Datenbank von vorherig gewonnenen Sequenzinformationen

Quellenverzeichnis

- Yunyue Zhu,Dennis Sasha.Statistical Monitoring of thousands of data streams in real time
- http://www.kx.com
- http://de.wikipedia.org
- http://en.wikipedia.org