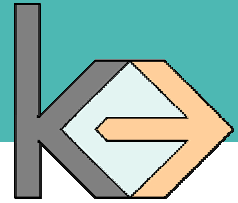
TECHNISCHE UNIVERSITÄT DARMSTADT



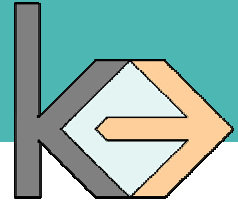
Signature-Based Methods for Data Streams

Corrina Cortes & Daryl Pregibon

Stefan Steger

10.01.2007

TUD



Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Einleitung

Features

- Scalars & Continuous Distribution
- Categorical Distribution & Item Sets
- Inferred (Statistical) Features
- Main Effect vs Interactions

Initializing & Updating Signatures

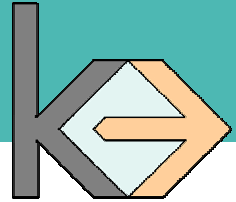
- Initializing
- Updating

Applications

- Customized Fraud Detection

Ausblick

Quellenangabe



Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Transaction Data Stream (TDS):

- Kontinuierlicher Datenfluss mit Aufzeichnungen von Transaktionen
  - Großes Volumen von einfachen Daten
  - Beispiel: Börsenhandel, Kredit Karten

Data Stream ⇔ Data Set

- kontinuierlich ⇔ statisch

Was ist eine Signatur?

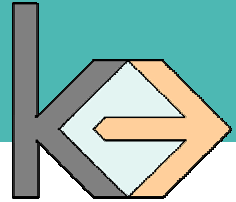
- erfasst das typische Verhalten von Benutzern

Wieso der Begriff Signatur?

- spiegelt den personifizierten Charakter der Daten wieder

Signature Processing

- event driven ⇔ time driven



Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Wieso Signaturen?

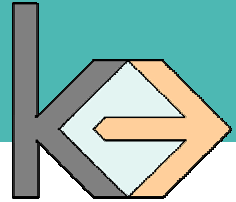
- Real-time characterization of users

Fraud Detection Methods

- Profil-based
  - Library mit Attack Profilen
  - Signaturen werden mit diesen Profilen verglichen
- Anomaly Detection Methods
  - Signatur ist selber Basis für Vergleich
  - Abweichung zeigt evt. Betrugsversuch an

Andere Anwendungsgebiete

- Marketing Analysen
  - Usage-based profiling
  - Signatur ist Grundlage für behavioral clustering



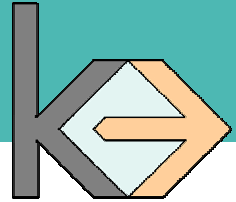
Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Domäne: Telekommunikation (AT&T)

- (International) Telecommunication Data Stream
- Call Detail Records (CDR)
- (International) Calling Signatur

Größenordnung

- Ende 1990 (international):
  - 5 Millionen Aufzeichnungen pro Tag
  - 12 Millionen Anschlüsse
- 2001 (wireless):
  - 80 Millionen Aufzeichnungen pro Tag
  - 15 Millionen Anschlüsse
- In Zukunft
  - 300 Millionen Aufzeichnungen pro Tag
  - 100 Millionen Anschlüsse



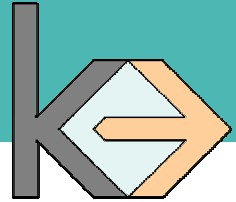
Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Welche Felder der Records (Transactions) werden verwendet?

- TDS enthält viele „Rohdaten“, aber nur einige sind relevant
  - Datum, Zeit, Dauer, Ursprung und Ziel des Anrufs, Bezahler des Anrufs

Wozu brauchen wir die Informationen aus den Records?

- Wollen wissen wie Kunden den Service nutzen - wann, wo und warum
- geändertes Verhalten erkennen → evt. Fraud?
- geringe Nutzung → anderer Anbieter?
- neue auf den Kunden abgestimmte Services entwickeln



Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Scalars

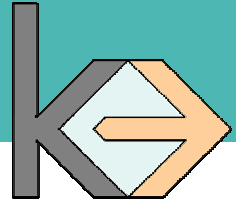
Beispiel: Anzahl Anrufe zu einem bestimmten Anschluß
Signature Processing

- period-by-period (time-driven)
  - Anzahl Anrufe pro Zeitperiode
- call-by-call (event-driven)
  - zusätzliches feld mit datum + zeit des letzten anrufs
  - daraus kann call rate abgeleitet werden

Continuous Distributions

Skalar nicht immer günstig

- Wenn sehr verzerrt
  - Wertebereich des Features in nichtüberlappende Behälter (bins) aufteilen
  - Danach den Wert quantisieren, in den jeweiligen Behälter tun und zählen
- Beispiel: Anrufdauer (call duration)
  - 0-100s-200s-8m-16m-32m-64m-128m-∞



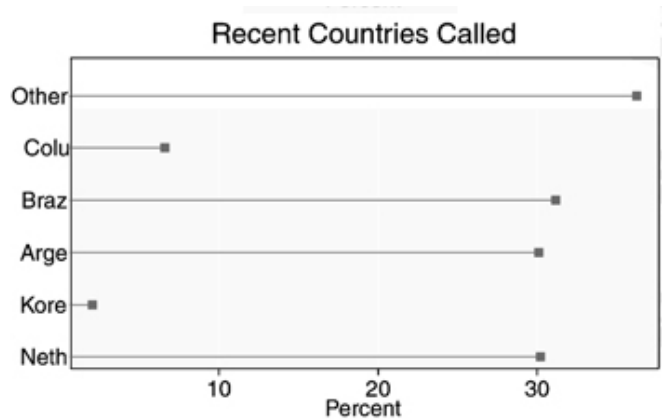
Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

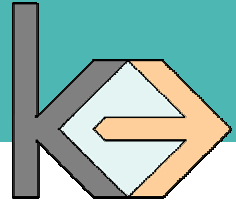
Categorical Distributions

- Behälter sind schon Teil der Definition der Variabeln
  - z.B.: Wochentag, Region der welt

Item Sets

- Categorical Variables mit vielen Werten
  - z.B.: Liste Internationale Rufnummern → ca. 100 Millionen Behälter
- Wollen Ziel von Anrufen zurückverfolgen
  - Liste ist zu groß um sie bei jedem account mitzuführen,
  - nur ein kleines Subset wird überhaupt angerufen
  - die besten k–Länder verwendet und eine Kategorie Rest.
- Behälter können sich mit der Zeit ändern



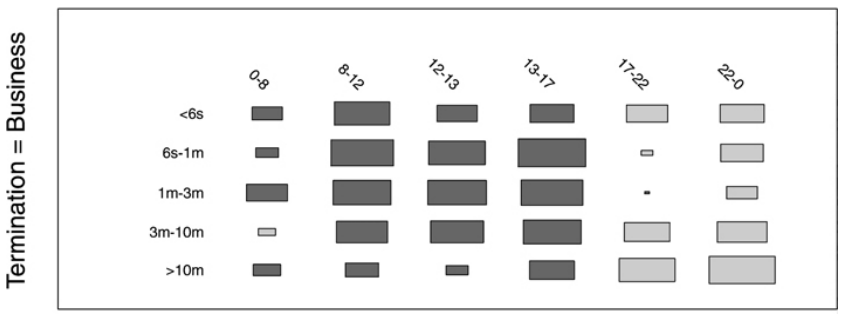


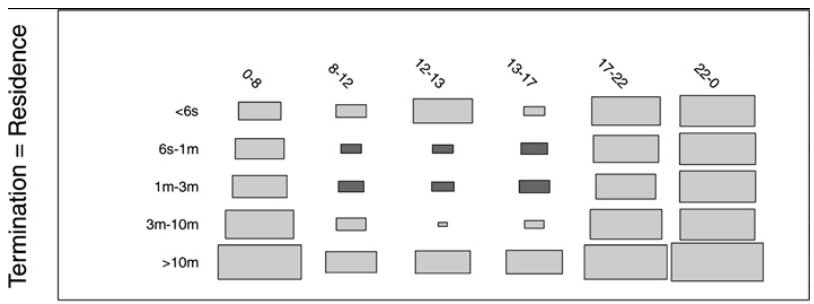
Inferred (Statistical) Features

Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

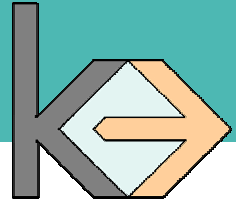
Beispiel: Bizocity - geschäftlicher oder privater Anschluß

- im CDR gibt es kein Feld dafür. Was also machen?
- man nimmt jeweils 1 Million bekannte geschäftliche und private Anschlüsse
- vergleicht **date, time, duration** and **type of destination**
- konstruieren jeweils einen Predictor Vector der Anrufe sortiert
- alle Vektoren werden mit Label (geschäftlich / privat) versehen
- Benutzen Regression Model um die Parameter der Scoring Funktion vorherzusagen (coefficients)





Coeffients of a logistic regresion model
business-like behavior (dark) or residence-like behavior (light).



Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Main Effects versus Interactions

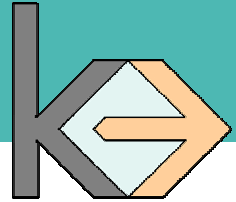
- Signaturen können sehr groß werden
  - Signaturgröße: 100 Million Accounts * 500-600 bytes ➔ 46-56 GB

Main Effect (voneiander möglichst unabhängige Features)

Interactions (Verknüpfungen zwischen Features)

- Nachteil: explosionsartige Anstieg der Signaturgröße
- Beispiel: International Calling Signature
  - Time of day ⇔ weekend / weekday
  - Time of day ⇔ day of week ist schon zu groß,
    120 integer werte mehr → Signaturgröße steigt um 180 GB an
- ist die statistische Sicherheit der extra Zellen gewährleistet?
- enthalten die neuen Zellen tatsächlich auch neue Informationen?
  - Gibt es wirklich mehr Betrugsfälle am Mittwochnachmittag?

- ➔ tradeOff zwischen zu erwartenden Zugewinn durch Interaktionen und Speicherplatz



Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Wieso müssen Signaturen Initialisiert werden?

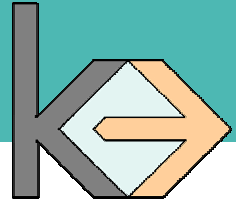
- für alle Anschlüsse müssen Signaturen vorhanden sein
- es entstehen ständig neue Anschlüsse – bis zu 50 000 täglich
- es kann Wochen dauern bis eine verlässliche Signatur entstanden ist

Wie wird Initialisiert?

- es werden Äquivalenzklassen von Anschlüssen gebildet + Evolution der Signaturen werden verfolgt
- neuer Anschluß → identifizieren der Äquivalenzklasse und initialisieren des Anschlusses mit der Durchschnittsignatur
- Ist äquivalent zu einer nearest neighbor classification

Wie finde ich die Aquivalenzklasse?

- man untersucht die ersten Anrufe – es reichen die ersten beiden
- timestamps → Vorhersage von calling rate & calling duration
- timestamps → Vorhersage von bizocity
- bizocity → Vorhersage von day of week und time of day
- destination → Vorhersage von region of world (top-k item list wird initialisiert)
- Ist alles nicht sehr präzise, reicht aber dennoch



Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Outliers

- falls neuer Record zu sehr von Signatur abweicht gibt es kein Update
  - könnte ein Betrugsversuch sein

Event-driven vs Time-driven

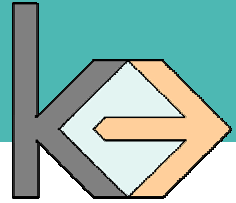
Ablauf des Updates

1. Signatur von Disk / Memory lesen
2. Werte in der Signatur verändern
3. Signatur zurück auf Disk / Memory schreiben

Nachteile und Vorteile der beiden Ansätze:

- event-driven
  - sehr zeitnah
  - I/O Anforderungen können bei größeren Datenbanken beachtlich sein.
  - Variabeln wie weekly calling rate sind schwerer vorherzusagen
- time-driven
  - weniger I/O gebunden
  - temporärer Speicherbedarf kann sehr groß werden
  - meisten tägliches time-driven updateing → brauchen disk-space für die Daten eines Tages

- Fraud-Detection-Systeme erfordern event-driven updating



Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Updating Algorithm

S_t : Signature zum Zeitpunkt t
R : Record / Set of Records
$\Delta = \Delta(R)$: transformiertes R
ε : Grenzewert für Outliers
Θ : Heuristic blending factor

$$S_{t+1} = \begin{cases} S_t & if\ |S_t - \Delta| > \varepsilon \\ \Theta S_t + (1-\Theta)\Delta & otherwise \end{cases}$$

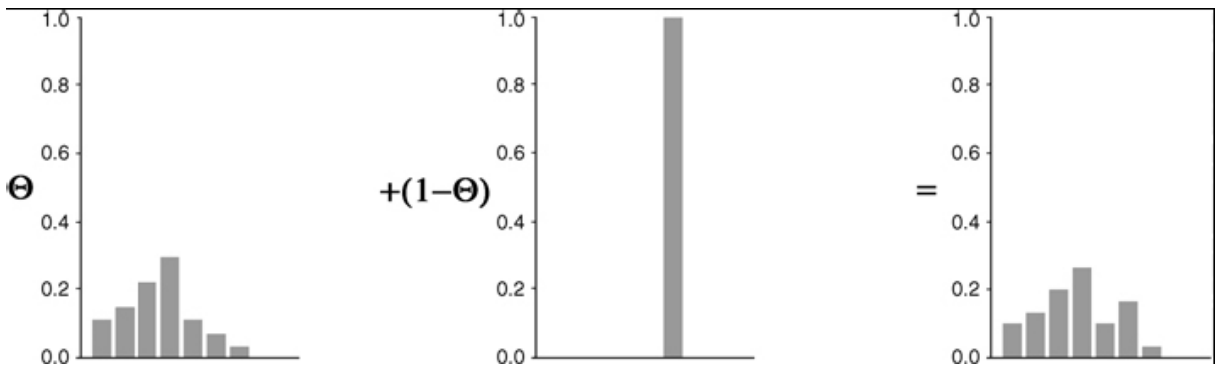
$0 < \varepsilon < \infty$

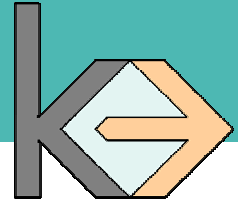
$0 \le \Theta \le 1$

Heuristic blending factor

time-driven
- ist Θ konstant
- Wert hängt vom gewünschten Zeitfenster ab
  - 0,85 → 30 Tage; 0,5 → 7 Tage

event-driven
- ist Θ Funktion der record interrarival time
  → Zeitfenster ist konstant
- ist $\Theta = \Theta(\text{calling rate})$
  → Bedeutung eines neuen Records hängt vom Benutzerverhalten ab





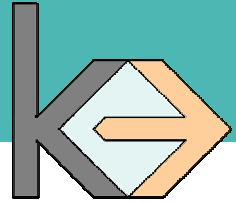
Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Updating Item Sets

- Item Sets werden etwas anders aktualisiert
  - da die Item Liste sich mit der Zeit ändert

Item Set Updating Algorithm

- wenn der Record ein Item aus der Item Liste enthält wird diese einfach aktualisiert
- Ansonsten wird eine Zufallszahl generiert
- Die top-k-Liste enthält einen „Abschußkandidaten“
  - Hängt von der Häufigkeit und Neuheit ab
- Zufallszahl wird mit der Wahrscheinlichkeit des Abschußkandidaten verglichen
  - Wenn Zufallszahl größer wird neues Item in die Liste genommen, altes fliegt raus



- Ziel: Betrugsversuche entdecken, möglichst wenige false negatives
- Kombination von beiden Ansätzen

Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Anomaly Detection Methods

- Ändert sich das Verhalten stark?
- Verdächtiges Verhalten kann für einen bestimmten Account normal sein
- Mißt wie ungewöhnlich ein Anruf ist

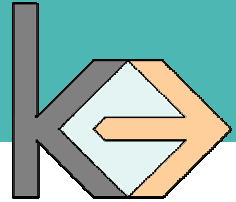
Profile Based

- Generic Fraudster Signatur
  - Daten dazu kommen von fraud investigatores
- Anruf wird sowohl mit der Generic Fraudster Signatur als auch der Account Signatur verglichen (fraudiness)

$$\text{fraudiness}(\text{call}) = \frac{\text{prob}(\text{call} \mid \text{customer signature})}{\text{prob}(\text{call} \mid \text{fraudster signature})}$$

- Charakterisiert einen ungewöhnlichen Anruf

- System ist aber anfällig für schleichende Unterwanderung
  - Wird allerdings vom Kunden entdeckt



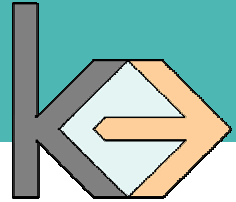
Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Fazit

- Die vorgestellten Methoden können auch woanders verwendet werden
  - Finanzwesen, Handel und Gesundheitswesen
- Sind aber nicht unumstritten
  - Datenschutz, Marketing Analysen
  - Einmal verdächtig – immer verdächtig?
- Man könnte noch Graphenalgos verwenden
  - Erkennen „calling circles“

Ausblick

- Hauptaugenmerk liegt aber auf noch größeren Datenmengen
  - Hancock
    - Ermöglicht high level programming of signature methods
    - www.research.att.com/~kfisher/hangcock/



Einleitung
Features
Initializing & Updating Signatures
Application
Ausblick

Quellen:

Cortes C.; Pregibon D. 2001 Signature-Based Methods for Data Streams

www.wikipedia.de

Fragen?