Sequential Pattern Analysis und Markov-Modelle.

Christian Weiß

Institut für Angewandte Mathematik und Statistitik

Universität Würzburg

Sequential Pattern Analysis

Historische Aspekte

Data Mining als Teildisziplin des

KDD
(Knowledge Discovery from Databases)

von großer Popularität.

Rasante Entwicklung begünstigt durch

- Möglichkeit, immer preiswerter riesige Mengen von Daten abzuspeichern

- Möglichkeit, mittels immer schnellerer Computer diese Datenmassen zu analysieren.

Ein Zweig des Data Mining,

Sequential Pattern Analysis,

entstand 1995, motiviert durch Alarmsequenzen in Telekommunikationsnetzwerken.

Problem: In Kontrollzentrum gehen unzählige, meist unbegründetete, Fehlermeldungen ein.

Ziel: Gibt es bestimmte Muster solcher Fehlermeldungen, auf die, innerhalb einer gewissen Zeitspanne, dann tatsächlich der Ausfall einer Netzkomponente folgt?

Pionierarbeit leisteten dabei

- **Agrawal, R., Srikant, R.**: *Mining sequential patterns*. 11^{th} Int. Conf. on Data Engg. (ICDE '95), Taipeh, Taiwan, pp. 3-14, 1995.

- **Mannila, H., Toivonen, H., Verkamo, A.I.**: *Discovering frequent episodes in sequences*. Proc. of the 1^{st} Int. Conf. on Knowl. Disc. and Data Mining (KDD '95), pp. 210-215, 1995.

- **Mannila, H., Toivonen, H., Verkamo, A.I.**: *Discovery of frequent episodes in event sequences*. Data Mining and Knowledge Discovery I, pp. 259-289, 1997.

Sequential Pattern Analysis

Typische Verfahren

Sequential Pattern Analysis – Typische Verfahren

Geg.: Kategoriale Zeitreihe $(X_t)_{\mathcal{T}}$,
wobei $\mathcal{T}$ abzählbar oder $\mathcal{T} = \{0, ..., T\}$ mit $T \in \mathbb{N}$.

$(X_t)_{\mathcal{T}}$ heißt *event sequence*.

Wertebereich der Zeitreihe $(X_t)_{\mathcal{T}}$:
Endliches Alphabet $\mathcal{V} := \{1, ..., q+1\}$ (*events*).

Ziel: Auffinden sog. *patterns* oder *episodes*:
Abschnitte $(X_{t-k}, ..., X_t) \equiv \alpha$, $\alpha \in \mathcal{V}^{k+1}$.

Sequential Pattern Analysis – Typische Verfahren

Im Allgemeinen komplexere Syntax erlaubt:

- *wildcard* '*' als Platzhalter,
- *parallel subepisode* $[i_1, ..., i_r]$ als logisches ODER.

Bsp.: Alphabet $\mathcal{V} := \{1, 2, 3\}$,
Muster $\alpha := (1, *, [2, 3])$ entspricht Menge
$\{(1, 1, 2, 3), (1, 2, 2, 3), (1, 3, 2, 3), (1, 1, 3, 2), (1, 2, 3, 2), (1, 3, 3, 2)\}$.

Sequential Pattern Analysis – Typische Verfahren

Typischer Aufbau von Algorithmen:

Schritt k:
Gegeben: Menge von *Kandidaten* der Länge k.
Menge enthält alle *häufigen* Muster der Länge k.

1. Finde alle häufigen Muster der Länge k.

2. Konstruiere Menge von Kandidaten der Länge $k+1$.

Apriori-Prinzip:
Muster der Länge $k+1$ *häufig* im Datensatz,
wenn *alle* Teilmuster der Länge k häufig.

Sequential Pattern Analysis – Typische Verfahren

Praktische Durchführung:

- Fenster fixer Länge $w+1$ gleitet über Daten,
- Inhalt des Fensters wird analysiert.
  $\Rightarrow$ Muster maximal der Länge $w+1$
- Erzeugen von *Regeln* der Art

$$\alpha \quad \Rightarrow \quad \beta, \qquad \text{mit } \beta = (\alpha, i_1, ..., i_r).$$

- Bewerten der gefundenen Regeln $\rightarrow$

Sequential Pattern Analysis – Typische Verfahren

Bewertung und Filterung der Regeln ($\rightarrow$ Data Mining zweiter Ordnung):

Zuordnung eines *confidence value* zu jeder Regel:

$$conf(\alpha \Rightarrow \beta) \ := \ \frac{N(\beta)}{N(\alpha)},$$

wobei $N(\gamma)$ die absolute Häufigkeit des Musters γ in der Zeitreihe.

Regel $\alpha \Rightarrow \beta$ *interessant*, wenn $conf(\alpha \Rightarrow \beta) \ \geq$ Schranke.

Sequential Pattern Analysis – Typische Verfahren

Sei nun $\beta \Rightarrow \gamma$ eine solche Regel und ihr confidence value berechnet.

Übliche Interpretation (Mannila et al. (1997)):

The confidence can be interpreted as the conditional probability of the whole of γ occuring in a window, given that β occurs in it.

... und das ohne Zugrundelegung irgendeines stochastischen Modells!

Fragestellung:

Gibt es stochastisches Modell,

dass solche Vorgehensweise und Interpretation rechtfertigt?

Dazu:

Welche Annahmen werden implizit gemacht?

Sequential Pattern Analysis

Implizite Annahmen

Sequential Pattern Analysis – Implizite Annahmen

X_t sei aktuellster Wert im *sliding window*.

Fenster hat die Länge $w + 1$

$\Rightarrow$ Impl. Annahme: X_t von max. w Vorgängern beeinflusst:

$$P(X_t = x_t \mid \mathcal{X}_{t-1}) \equiv P(X_t = x_t \mid X_{t-1} = x_{t-1}, ..., X_{t-w} = x_{t-w})$$

$\Rightarrow$ Impl. Annahme: **Markov-Modells der Ordnung** w

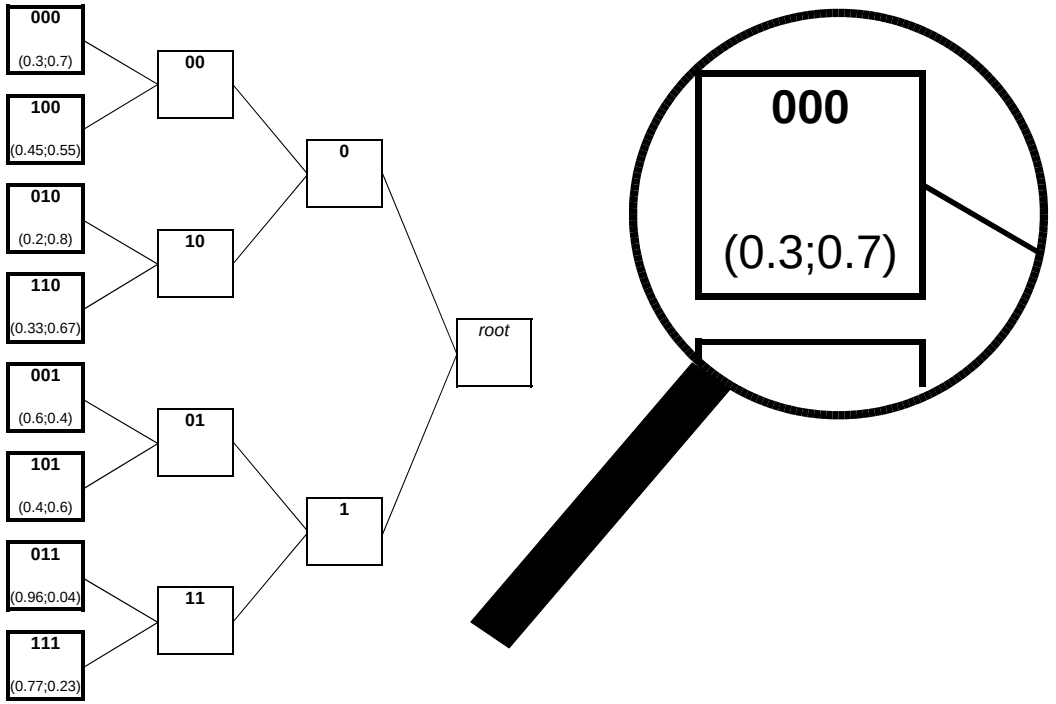
Ferner: Zeitpunkt des Auftretens von Mustern

ohne Einfluss auf confidence value $conf(\alpha \Rightarrow \beta) := \frac{N(\beta)}{N(\alpha)}$

$\Rightarrow$ **Homogenes Markov-Modell der Ordnung** w.

Homogene Markov-Modelle

Probabilistic Suffix Tree (PST) am Beispiel eines binären Markov-Modells der Ordnung 3:



Homogene Markov-Modelle

Sei $(X_t)_{\mathcal{T}}$ Zeitreihe über Alphabet $\mathcal{V}$

$(X_t)_{\mathcal{T}}$ genüge homogenem Markov-Modell der Ordnung w.

Log-Likelihood-Funktion von $(X_t)_{\mathcal{T}}$:

$$\ell_{\boldsymbol{X}}(\mathcal{P}) \;:=\; \ln I \;+ \sum_{(i_0,i_{-1},\ldots,i_{-w})\in\mathcal{V}^{w+1}} N(i_0,i_{-1},\ldots,i_{-w}) \cdot \ln p(i_0|i_{-1},\ldots,i_{-w}),$$

wobei initialer Term

$$I \;:=\; I(x_{w-1},\ldots,x_0) \;:=\; P(X_{w-1}=x_{w-1},\ldots,X_0=x_0).$$

Homogene Markov-Modelle

Wegen $p(q+1|i_{-1},...,i_{-w}) \;=\; 1-\sum_{i_0=1}^{q} p(i_0|i_{-1},...,i_{-w})$

erhalte partielle Ableitungen

$$\frac{\partial}{\partial p(i_0|i_{-1},...,i_{-w})}\,\ell_{\boldsymbol{X}}(\mathcal{P}) \;=\; \frac{N(i_0,i_{-1},...,i_{-k})}{p(i_0|i_{-1},...,i_{-w})} \;-\; \frac{N(q+1,i_{-1},...,i_{-w})}{p(q+1|i_{-1},...,i_{-w})}.$$

Somit **ML-Schätzer** für $p(i_0|i_{-1},...,i_{-w})$:

$$\widehat{p}(i_0|i_{-1},...,i_{-w}) \;:=\; \frac{N(i_0,i_{-1},...,i_{-w})}{N(\bullet,i_{-1},...,i_{-w})}.$$

Entspricht exakt dem **confidence value** für die Regel

$$(i_{-w},...,i_{-1},*) \;\Rightarrow\; (i_{-w},...,i_{-1};i_0).$$

Sequential Pattern Analysis – Implizite Annahmen

Annahme eines w-Markov-Modells $\Rightarrow$

Für Regeln *voller* Länge gilt:

confidence value = ML-Schätzer für zug. bedingte Wahrscheinlichkeit.

Allerdings auch andere Typen von Regeln:

Kürzere Regeln $(i_{-k}, ..., i_{-1}, *) \Rightarrow (i_{-k}, ..., i_{-1}; i_0), \quad k < w.$

Dann implizite Annahme für alle $i_0 \in \mathcal{V}$,
dass noch frühere Werte ohne Einfluss:

$$P(X_t = i_0 \mid ..., X_{t-w} = i_{-w}) = P(X_t = i_0 \mid ..., X_{t-k} = i_{-k}).$$

Sequential Pattern Analysis – Implizite Annahmen

Log-Likelihood-Funktion: Term

$$\sum_{i_0, i_{-(k+1)}, ..., i_{-w} \in \mathcal{V}} N(i_0, i_{-1}, ..., i_{-w}) \cdot \ln p(i_0 | i_{-1}, ..., i_{-w})$$

vereinfacht sich zu

$$\sum_{i_0=1}^{q+1} N(i_0, i_{-1}, ..., i_{-k}, \bullet, ..., \bullet) \cdot \ln p(i_0 | i_{-1}, ..., i_{-k}).$$

Ergo: **ML-Schätzer** für $p(i_0 | i_{-1}, ..., i_{-k})$ ist

$$\widehat{p}(i_0 | i_{-1}, ..., i_{-k}) \ := \ \frac{N(i_0, i_{-1}, ..., i_{-k})}{N(\bullet, i_{-1}, ..., i_{-k})}.$$

Dies erneut gleich dem **confidence value** der Regel

$$(i_{-k}, ..., i_{-1}, *) \ \Rightarrow \ (i_{-k}, ..., i_{-1}; i_0).$$

Sequential Pattern Analysis – Implizite Annahmen

Analoges Vorgehen auch bei Regeln

$$(i_{-k}, ..., i_{-1}, *, ..., *) \;\Rightarrow\; (i_{-k}, ..., i_{-1}; i_0, ..., i_r)$$

mit **Kopf der Länge** $r + 1 > 1$, wobei $r + k \leq w$.

Annahme: Frühere Werte als i_{-k} bedeutungslos

$\Rightarrow$ Faktorisiere bedingte Wahrscheinlichkeit

$$p(i_r, ..., i_0 | i_{-1}, ..., i_{-k}) \;=\; \prod_{j=0}^{r} p(i_j | i_{j-1}, ..., i_{-k}).$$

Sequential Pattern Analysis – Implizite Annahmen

Schätze Faktoren $p(i_j|i_{j-1},...,i_{-k})$ analog oben (mit entsprechenden Einschränkungen)

$\Rightarrow$ **ML-Schätzer** $\widehat{p}(i_r,...,i_0|i_{-1},...,i_{-k}) =$

$$\prod_{j=0}^{r} \frac{N(i_j,...,i_{-k})}{N(\bullet,i_{j-1},...,i_{-k})} = \frac{N(i_r,...,i_{-k})}{N(\bullet,...,\bullet,i_{-1},...,i_{-k})}.$$

Entspricht erneut **confidence value** der Regel

$$(i_{-k},...,i_{-1},*,...,*) \Rightarrow (i_{-k},...,i_{-1};i_0,...,i_r).$$

Sequential Pattern Analysis – Implizite Annahmen

Regeln, deren **Kopf von komplexer Syntax**:

$$(i_{-k}, ..., i_{-1}, *, ..., *) \;\Rightarrow\; (i_{-k}, ..., i_{-1}; \boldsymbol{f}(i_0, ..., i_r)),$$

wobei $\boldsymbol{f}(i_0, ..., i_r)$ abkürzende Schreibweise für Ausdruck in $i_0, ..., i_r$ mit *wildcards* und/oder *parallel operator*.

Verwende, dass

$$p(\boldsymbol{f}(i_r, ..., i_0)|i_{-1}, ..., i_{-k})$$

$$= \sum_{(j_r,...,j_0)\in \boldsymbol{f}(i_0,...,i_r)} P(X_{t+r} = j_r \, ..., X_t = j_0 \mid ..., X_{t-k} = i_{-k})$$

... $\Rightarrow$ confidence value = ML-Schätzer.

Sequential Pattern Analysis – Implizite Annahmen

Schwierigster Fall: Komplexe Syntax in der Basis der Regel, z.B. der Art

'If B or C, then A with probability p.'

Dies beinhaltet die implizite Annahme

$$p \;:=\; P(A \mid B \vee C) \;=\; P(A \mid B) \;=\; P(A \mid C).$$

Sequential Pattern Analysis – Implizite Annahmen

Bei Regeln des Types

$$(X_{t-1}, ..., X_{t-k}) \in \boldsymbol{f}(i_{-1}, ..., i_{-k}) \quad \Rightarrow \quad X_t = i_0,$$

implizite Annahme, dass

$$p(i_0|j_{-1}, ..., j_{-k}) \;\equiv\; p(i_0|\boldsymbol{f}(i_{-1}, ..., i_{-k}))$$

für alle $(j_{-1}, ..., j_{-k}) \in \boldsymbol{f}(i_{-1}, ..., i_{-k})$ und $i_0 \in \mathcal{V}$.

Entsprechend erhält man als ML-Schätzer

$$\widehat{p}(i_0|\boldsymbol{f}(i_{-1}, ..., i_{-k})) \;:=\; \frac{N(i_0, \boldsymbol{f}(i_{-1}, ..., i_{-k})}{N(\bullet, \boldsymbol{f}(i_{-1}, ..., i_{-k}))}.$$

Sequential Pattern Analysis – Zusammenfassung

Interpretiere **confidence value** einer sequentiellen Regel als **ML-Schätzer** für zug. **bedingte Wahrscheinlichkeit**, wenn

- Annahme, dass Daten beruhen auf **homogenem Markov-Modell der Ordnung** w,

  und

- Annahme obiger Einschränkungen (wobei diese ohnehin bereits implizit gemacht).

  $\Rightarrow$ **Variable Length Markov Model (VLMM)**

Sequential Pattern Analysis

Modell

Sequential Pattern Analysis – Modell

Homogenes *variable length Markov model (VLMM) von Ordnung* k:

- $(X_t)_{\mathcal{T}}$ ist homogener k-Markov-Prozess

- es gibt Tupel $(i_{-1}, ...i_{-s})$ mit $s \leq k$, so dass

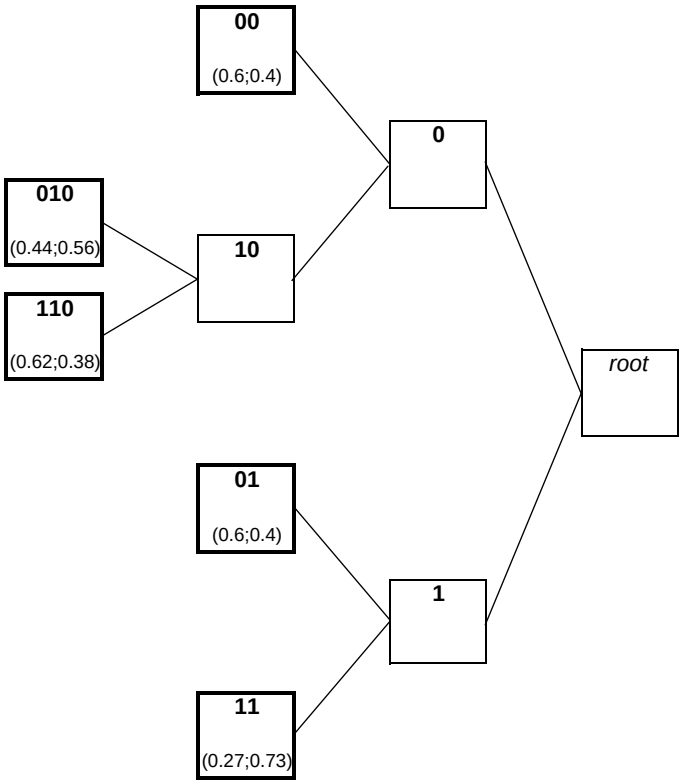
  $P(X_t = i_0 \mid X_{t-1} = i_{-1}, ..., X_{t-k} = i_{-k}) = p(i_0|i_{-1}, ..., i_{-s})$

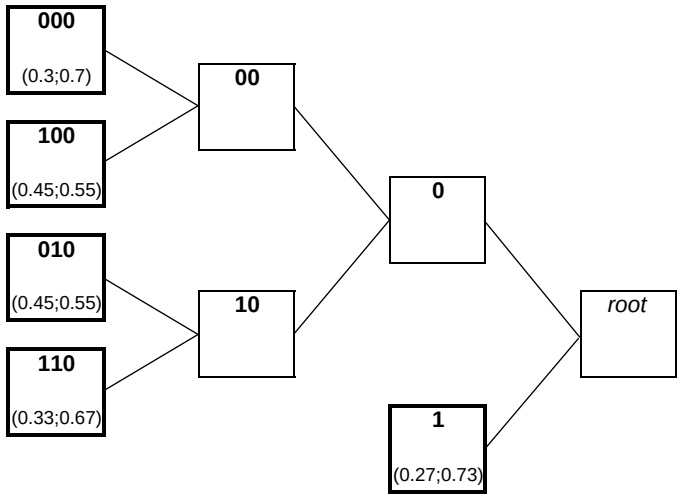
  für alle $(i_{-s-1}, ..., i_{-k}) \in \mathcal{V}^{k-s}$ und $i_0 \in \mathcal{V}$.

Vorteil gegenüber gewöhnlichem Markov-Modell: Weniger zu schätzende Parameter bei ähnlicher Flexibilität.

Sequential Pattern Analysis – Modell

Beispiele für binäre VLMM der Ordnung 3, dargestellt als PST:





Sequential Pattern Analysis – Modell

Wichtige Artikel zum VLMM:

- **Rissanen, J.**: *A universal data compression system*.
  IEEE Trans. on Inf. Theory 29 (5), pp. 656-664, 1983.

- **Ron, D., Singer, Y., Tishby, N.**: *The power of amnesia: Learning probabilistic automata with variable memory length*.
  Machine Learning 25, pp. 117-149, 1996.

- **Bühlmann, P., Wyner, A.J.**: *Variable length Markov chains.*
  Annals of Statistics 27, pp. 480-513, 1999.

Sequential Pattern Analysis

Zukünftige Strategie

Sequential Pattern Analysis – Zukünftige Strategie

Nachdem gezeigt wurde, dass ein VLMM ohnehin bereits implizit angenommen wurde, empfiehlt sich zukünftig folgendes Vorgehen:

1. Anpassung eines VLMM an die Daten bzw. Konstruktion eines geeigneten PST.

2. Erzeugung von Regeln unter Berücksichtigung des Modells.

Sequential Pattern Analysis – Zukünftige Strategie

Vorteile:

zu 1.: In der oben erwähnten Literatur gibt es Verfahren und Tests zur Anpassung eines geeigneten VLMM
$\Rightarrow$ **Verbesserung** bisheriger Verfahren.

zu 2.: Da Regeln nur für die Blätter des PST erzeugt werden, führt dies zu einer sinnvollen Reduktion der Zahl der erzeugten Regeln
$\Rightarrow$ **Optimierung** bisheriger Verfahren.

Vielen Dank für Ihr Interesse!