Learning Multiple Evolutionary Pathways from Cross-Selectional Data

Niko Beerenwinkel Jörg Rahnenführer Martin Däumer
Daniel Hoffmann Rolf Kaiser Joachim Selbig
Thomas Lengauer

vorgestellt und zusammengefasst durch: *Sebastian Zeidler* im Rahmen des Data-Mining Seminars im WS05/06

Inhaltsverzeichnis

1 Definitionen eines Mutationsbaumes

$\ell :=$ Menge der Mutationsereignisse
$x_i := x_{i1..i\ell}$ repräsentiert diese Ereignisse
$X = (x_{ij})_{\substack{1 \le i \le N \\ 1 \le j \le \ell}}$
$\Omega = 2^{\{1,....,\ell\}}$ ist die Menge aller möglichen Kombinationen
im vorliegenden Beispiel ist $N = 367$ und $\ell = 7$
Somit ist der Baum: $\tau = (V, E, r, p)$ wobei $V =$ Events, $E =$gewichtete Kanten, $r =$Startknoten, $p : E \to [0, 1]$ Wahrscheinlichkeiten
für alle Kanten gilt: $e = (j_1, j_2) \in E : p(e) = Pr(j_2|j_1)$

2 Baumaufbau und Berrechnung der Likelihood

$w(j_1, j_2) = logPr(j_1, j_2) - log(Pr(j_1) + Pr(j_2)) - logPr(j_2)$ wurde durch Desper gezeigt kann der Baum aufgebaut werden
$L(x|\tau) = Pr(x|\tau)$ ist die Likelihood das ein Muster x durch einen Baum τ erzeugt wird
sind alle Mutationen x aus r erreichbar dann ist auch die Likelihood >0 sonst =0

3 Anwendung der Mutationsbäume für Mixtur Modelle

Die Mutationsbäume sind statisch und es können mehrere existieren, dies ist die Vorraussetzung für die Bildung eines Mixtur Modells
diese folgen einer diskreten multivariaten Verteilung $Y_1, ..., Y_k$
demnach sind die Bäume: $\tau_k = (V, E_k, r, p_k), k = 1, .., K$
Sei $\triangle_1, ...\triangle_k \in \{0, 1\}$ mit $Pr(\triangle_k = 1) = \alpha_k$
dann ist $\mathcal{M} = \sum_{k=1}^{K} \alpha_k \tau_k$ ein Modell mit
$\sum_{k=1}^{K} \alpha_k = 1$ und $\alpha_k \in [0, 1]$
somit ist: $Y = \sum_{k=1}^{K} \triangle_k Y_k$ ein $K-$Baum Mutations-Mixtur-Modell
daher ist die umdefinierte Likelihood: $L(x|\mathcal{M}) = \sum_{k=1}^{K} \alpha_k L(x|\tau_k)$
außerdem wird ein Sternmodell $k = 1$ eingeführt
hierbei gilt: $p(e) = \beta$ für alle $e \in \tau_1$

4 EM ähnlicher Lern Algorithmus

Wir bauen nun einen Algorithmus für die Maximierung der Likelihood unter dem Aspekt, dass wir einen maximal wahrscheinliches Modell suche
Wir definieren das Mixtur-Modell: $\sum_{i=1}^{N} log \sum_{k=1}^{K} \alpha_k L(x_i|\tau_k)$
wobei $\alpha_1 ... \alpha_k$ die unterschiedlichen Baum Komponenten sind
sind alle x_i unabhängig ist die Antwort auf die Modellkomponenten k folgende $\gamma_{ik} = Pr(\triangle_k = 1|\mathcal{M}, x_i)$
Sei $N_k = \sum_{i=1}^{N} \gamma_{ik}$ die gewichtete Anzahl an Beispielen die durch τ_k erzeugt wurden
im Vergleich zum EM ist γ der E Schritt und die Berrechnung von $\mathcal{M}$ der M-Schritt
dann ist im nächsten Schritt bei gegebenem $\mathcal{M} = \sum_{k=1}^{K} \alpha_k \tau_k$
$\gamma_{ik} = \frac{\alpha_k L(x_i|\tau_k)}{\sum_{m=1}^{K} \alpha_m L(x_i|\tau_m)}$
außerdem ist auch in der nächsten Iterartion das Stern Modell neu zu berrechnen
$\beta = \frac{1}{\ell N_1} \sum_{j=1}^{\ell} \sum_{i=1}^{N} \gamma_{i1} x_{ij}$
Die Initialwerte werden vorher mittels (K-1)-Means geschätzt
es gilt also: $\gamma_{ik} = \begin{cases} \frac{1}{2}, & wenn\, x_i\, im\, Cluster\, k-1\, ist \\ \frac{1}{2(K-1)}, & sonst \end{cases}$

5 Ablaufplan

Initialisierung mittels (K-1)-Means und setzen der Anfangswerte für γ_{ik}
M-ähnlicher Schritt für das Update des Modells

- setze $N_k = \sum_{i=1}^{N} \gamma_{ik} \forall k = 1, ..., K$
- berrechne neues τ_1 als $\beta = \frac{1}{\ell N_1} \sum_{j=1}^{\ell} \sum_{i=1}^{N} \gamma_{i1} x_{ij}$
- Für k=2...K:
  - Berrechne alle Paare von Ereignissen $(j_1, j_2), 1 \leq j_1, j_2 \leq \ell$ und bestimmte
    $$p_k(j_1, j_2) = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_{ij_1} x_{ij_2}$$
  - berrechne das Maximum aller T_k mit den Gewichten w aus p_k
  - bestimme die Mixtur Parameter $\alpha_k = \frac{N_k}{N}$

E-Step Bestimme die neuen γ_{ik} Iteriere Schritte 2 und 3 bis ein Konvergenzkriterium erfüllt ist