Emerging Patterns und Supervised Learning

Anne-Laure Boulesteix

19 Februar 2004

Struktur

1. Problemstellung
2. Definition der Emerging Patterns
3. Emerging Patterns mit Bäumen finden
4. Simulationen
5. Anwendung für das Supervised Learning

1. Problemstellung

1. **Klassifikation mit Genexpressionsdaten**
   Binäre Response-Variable Y (Tumorart)
   p metrische Kovariaten $X_1, \ldots, X_p$
   (Genexpression)
   n Beobachtungen

   Ziel: Y anhand von $X_1, \ldots, X_p$ prädiktieren

   Problem: Zu viele Variablen ! $n \approx 100$, $p \approx 3000$

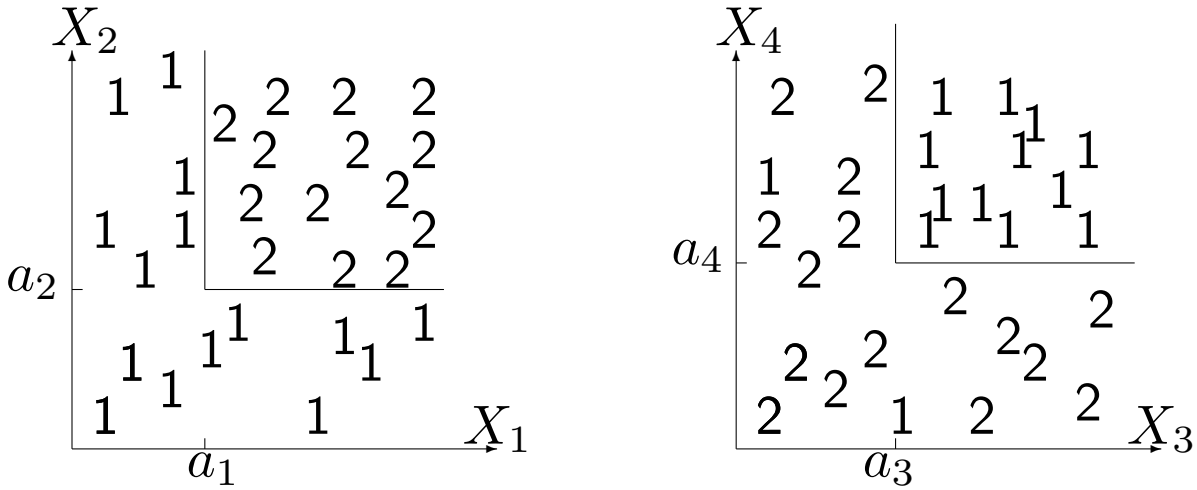
   Typischer Einsatz: Variablenselektion, um dann eine klassische Klassifikationsmethode anwenden zu können (kNN, LDA,...)

2. **Finden von Interaktionen**
   Interaktionen zwischen Variablen sind für die Biologen interessant.

Unser Ziel: Interaktionen finden und für die Klassifikation anwenden.

2. Definition der Emerging Patterns (1)

von Dong und Li (1999) eingeführt



Solche Interaktionen sind sowohl für biologische Forschung als auch für Supervised Learning interessant.

Formal: Ein EP ist ein Muster der Form

$$P = (X_{i_1} \leq a_{i_1}) \cap \cdots \cap (X_{i_m} \leq a_{i_m})$$

mit $|\log \frac{p(P|Y=1)}{p(P|Y=2)}|$ groß

Beschränkung: $m = 2$

3. EPs mit Klassifikationsbäumen finden

Idee: EP $\approx$ Blatt eines Klassifikationsbaums

Vorgehensweise:

1. Falls es sehr viele Variablen gibt, Variablenselektion durchführen

2. Baum mit der Tiefe 2 wachsen lassen

3. Für jedes Blatt testet man mit Fishers exaktem Test (zum Konf. p_S) die H_0-Hyp, dass die zweite Zerlegung klassifikationsrelevant ist. Wenn H_0 abgelehnt wird, eliminiert man das Blatt.

4. Für jedes ausgewählte Blatt P testet man mit Fishers exaktem Test (zum Konf. p_G) die H_0-Hyp $p(P|Y=1) = p(P|Y=2)$. Wenn H_0 abgelehnt wird, wählt man das Blatt als EP aus.

5. Die Variable, die die erste Zerlegung des Baums definiert, eliminieren. Zurück zu Schritt 2.

4. Simulationen

100 Zufallsmatrizen mit 100 Beobachtungen (20 EPs und 980 anderen Variablen).

Es kann sein, dass EPs per Zufall auftauchen, aber man ist nur an den 20 richtigen EPs interessiert.

Für jede Zufallsmatrix gilt:

- Hit rate = Anteil der richtigen EPs, die vom Algorithmus entdeckt wurden
- False alarm rate = Anteil der 'nicht-EP Paare', die als EP entdeckt wurden

Ergebnisse:

$$\text{Hit rate} \approx 50\%$$

$$\text{False Alarm Rate} \approx 0.001\%$$

5. Anwendung für das Supervised Learning

Learningsdatensatz $\mathcal{L}$, Testdatensatz $\mathcal{T}$

Vorgehensweise:

- Algorithmus auf $\mathcal{L}$ laufen lassen $\rightarrow m$ EPs werden gefunden
- Neue Variablen $Z_1, \ldots, Z_m$ definieren:

$$\begin{array}{rcll} Z_j & = & 1 & \text{falls } j\text{-tes EP erfüllt ist} \\ & = & 0 & \text{sonst} \end{array}$$

- neue Datenmatrizen $\mathbf{Z}_L$ und $\mathbf{Z}_T$ bestimmen
- Lineare Diskriminanzanalyse mit Z_L und Z_T durchführen
- Fehlerrate berechnen

Die ganze Prozedur wird mit 50 zufälligen Zerlegungen wiederholt.

Diskussion

- Die Klassifikationsergebnisse sind so gut wie mit den klassischen Diskriminanzmethoden
- Die Interaktionseffekte werden schnell und relativ effizient entdeckt

Verbesserungsvorschläge und Ausblick:

- Statistische Relevanz von EPs
- Mehrkategoriale Response-Variablen
- Interaktionen mit mehr als 2 Variablen.

Literatur:

Dong,G., Li,J. (1999), Efficient Mining of Emerging Patterns: Discovering Trends and Differences, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining ACM Press, San Diego, CA, pp.43 – 52