Zusammenfassung des Artikels: "Using a Mixture of Probilistic Decision Trees for Direct Prediction of Protein Function"

Umar Syde and Golan Yona
Vortragender: Thomas Schmutzer

29. März 2006

1 Grundlagen & Decision Trees

Attributkategorien:

seqeunce features
database features
predicted features

traditionelles Lernen:

greedy Attribut Selektion
Entropie $= -\sum_{j=1}^{c} p_j \; log_2 p_j$
$Gain(S, A) = Entropie(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$

2 Optimierung der Trainingsprozedur

missing values:

Verteilung auf Kindknoten entsprechend der prozentualen Aufteilung der übrigen

binary splitting:

binäre Baumstruktur gefordert
numerische Daten –> Spaltwert teilt Daten in zwei Bereiche
nominale Daten –> CART Algorithmus

Post-Pruning:

konstruierten Baum in bottom-up Strategie stutzten
bei Performanceeinbruch abbrechen

3 Mixture of probabilistic decision trees

Attributselektion:

$$Prob(A) = \frac{Gain(S, A)}{\sum_i Gain(S, A_i)}$$

Konstruktion des Mixture Modell:

besteht für eine Proteinfamilie aus jeweils 100 PDTs

Evaluierung der Entscheidungsbäume:

Jensen-Shannon Divergenz $D_\lambda^{JS}[p||q] = \lambda D^{KL}[p||r] + (1-\lambda)D^{KL}[q||r]$
mit D^{KL}als Entropiemaß für p bezüglich q und $r = \lambda p + (1-\lambda)q$

Verteilung der Daten:

zu wenig Positive im Datensatz –> uninformativ für Algorithmus
gemischte Entropie: $\frac{Entropie(S_g)+Entropie(S_u)}{2}$
besser für Attributselektion und Pruning auf unausgeglichenen Daten

Prinzip der " Minimal Description Length "

optimale Hypothese (PDT)

$$h^* = argmin[-P(h)P(D/h)] = argmin[-log_2P(h) - log_2P(D/h)]$$

mit Likelihood $P(D/h) = \prod_{i=1}^{n} P_T(x_i) = \prod_{i=1}^{n} \sum_{j \in leaves(T)} f_j(x_i) Prob_j(x_i)$ und
Komplexität (P(h) des Modell: Summe Binärdarstellungen $logN + logk_j$ aller Knoten
–> Vorteil für kleine Datensätze (keine Validierungsmenge für Post-Pruning notwendig)

4 Klassifikationstest mit Pfam und EC

Optimierung der Methoden zum Mixture Modell

–> enthält: binäres Splitting, gewichtete Entropie, Kreuzvalidierung aus 10 Mengen, mit JS-Divergenz optimiertes Post-Pruning und Evaluierung, Dipeptidinformationen und die probabilistic decision trees!
Effizienzsteigerung von 0.35 (C4.5) auf 0.81(MMPDTs)

Pfamklassifikationstest:

464744 Proteine aus SWISS-PROT und TrEMBL in Pfam-Familien eingeteilt
233 relevante Proteinklassen je in Lern- und Testdatensatz aufgeteilt
für jede Proteinfamilien MMPDTs(mixture modell of prob. decision tree) trainiert
vergleichende Analyse mit BLAST:
durchschnittliche Performance der PDTs in den Familien: 81%
durchschnittliche Performance BLAST(beste|gemittelte) in den Familien: 94%/|86%
Auswertung:
Test bescheinigt Potential und Gültigkeit der Strategie
mögliche Fehlerquelle: gemischte Entropie & aggresives Post-Pruning

EC-Klassifikationstest:

für 229 relevante EC-Proteinfamilien MMPDTs trainiert
durchschnittliche Performance der PDTs in den Familien: 71%
in 17 von 30 Fällen PDTs besser als der beste BLAST hit!!
gemittelte BLAST hits fast immer deutlich schlechter als PDTs!!
Auswertung:
rein funktionsbasierte Anaylse (falls Homologien in Sequenzen fehlen)
für BLAST ungenau
–> Funktionsähnlichkeit nicht zwingend mit Sequenzähnlichkeit korreliert!