Multi-label Lazy Associative Classification

Darko Popovic

Seminar aus maschinellem Lernen

WS 2007/2008

Klassifizierung

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
12-30	mild	rain	high	false	yes

today	cool	sunny	normal	false	?
tomorrow	mild	sunny	normal	false	?

Klassifikatoren

- Binärer Klassifikator
- Multi-class Klassifikator

  → eine Klasse pro Instanz
- Multi-label Klassifikator

  → mehrere Label pro Instanz

Gliederung

- Motivation
- Grundlagen
- Multi-class Klassifikator
- Multi-label Klassifikator
- Experimente und Ergebnisse
- Fazit

Motivation

- Große Bedeutung der Multi-label Klassifizierung

  (z.B. Text-, Genklassifikation,…)
- Wenig erforscht!
- Herausforderungen:
  - „small disjuncts“
  - Abhängigkeiten zwischen Labels

Grundlagen

- Class association rule (CAR)

  Regel der Form: $F \xrightarrow{\sigma,\theta} l$

  F=Menge von Features, l=Label

- Support σ

  support(F→l) = P(F U l)

- Confidence θ

  confidence(F→l) = P(l|F)

Beispiel: CARs

Trainingsdaten:

ID	Label	Title	Actors
1	Drama/Mystery	Solaris	G. Clooney
2	Drama/Thriller	Syriana	G. Clooney, M. Damon
3	Crime/Thriller	Ocean's Eleven	G. Clooney, M. Damon
4	Crime/Thriller	Out of Sight	G. Clooney
5	Thriller	The Peacemaker	G. Clooney
6	Action	Desperado	Q. Tarantino

$\text{G. Clooney} \rightarrow \text{Drama}$

support (G. Clooney→Drama) = 2/6 =0.33

confidence (G. Clooney→Drama) = 2/5 =0.4

$\text{G. Clooney} \xrightarrow{0.33;0.4} \text{Drama}$

Alle CARs

$$\text{G. Clooney} \xrightarrow{0.33;0.4} \text{Drama}$$
$$\text{G. Clooney} \xrightarrow{0.17;0.2} \text{Mystery}$$
$$\text{G. Clooney} \xrightarrow{0.67;0.8} \text{Thriller}$$
$$\text{G. Clooney} \xrightarrow{0.33;0.4} \text{Crime}$$
$$\text{G. Clooney \& M. Damon} \xrightarrow{0.17;0.5} \text{Drama}$$
$$\text{G. Clooney \& M. Damon} \xrightarrow{0.33;1} \text{Thriller}$$
$$\text{G. Clooney \& M. Damon} \xrightarrow{0.17;0.5} \text{Crime}$$
$$\text{M. Damon} \xrightarrow{0.17;0.5} \text{Drama}$$
$$\text{M. Damon} \xrightarrow{0.33;1} \text{Thriller}$$
$$\text{M. Damon} \xrightarrow{0.17;0.5} \text{Crime}$$
$$\text{Q. Tarantino} \xrightarrow{0.17;1} \text{Action}$$

Multi-class Klassifikator

- CAR in Modell aufnehmen, falls:

  $\sigma \geq \sigma_{min}$ und $\theta \geq \theta_{min}$

- Modell mit $\sigma_{min} = 0.2$ und $\theta_{min} = 0.4$

$$\text{G. Clooney \& M. Damon} \xrightarrow{0.33;1} \text{Thriller}$$

$$\text{G. Clooney} \xrightarrow{0.67;0.8} \text{Thriller}$$

$$\text{G. Clooney} \xrightarrow{0.33;0.4} \text{Drama}$$

$$\text{G. Clooney} \xrightarrow{0.33;0.4} \text{Crime}$$

$$\text{M. Damon} \xrightarrow{0.33;1} \text{Thriller}$$

Multi-class Klassifikator

- Klassifiziere einem Film mit G. Clooney:

$$\text{G. Clooney} \xrightarrow{0.67;0.8} \text{Thriller} \quad s(\text{Thriller})=0{,}54$$

$$\text{G. Clooney} \xrightarrow{0.33;0.4} \text{Drama} \quad s(\text{Drama})=0{,}13$$

$$\text{G. Clooney} \xrightarrow{0.33;0.4} \text{Crime} \quad s(\text{Crime})=0{,}13$$

Thriller oder Drama oder Crime?

- Definiere Score-Funktion:

$$s(l_i) = \sum_{F \xrightarrow{\sigma,\theta} l_i \in \mathrm{M}} \sigma \times \theta$$

F = Teilmenge von Testinstanz

Multi-label Klassifikator (IEAC)

- Unabhängiger Klassifizierer für jedes Label
- Definiere Wahrscheinlichkeits-Funktion:

$$f\left(l_i\right) = \frac{s(l_i)}{max_l\, s(l)}$$

Weise Label l_i zu, falls $f\left(l_i\right) = \delta \geq \delta_{min}$

Beispiel: IEAC–small disjuncts

Trainingsdaten:

ID	Label	Title	Actors
1	Drama/Mystery	Solaris	G. Clooney
2	Drama/Thriller	Syriana	G. Clooney, M. Damon
3	Crime/Thriller	Ocean's Eleven	G. Clooney, M. Damon
4	Crime/Thriller	Out of Sight	G. Clooney
5	Thriller	The Peacemaker	G. Clooney
6	Action	Desperado	Q. Tarantino

Testinstanz:

7	?[Action]	From Tusk till Dawn	Q. Tarantino, M. Damon

Modell: $\sigma_{min} = 0.2;\ \theta_{min} = 0.67;\ \delta_{min} = 0.5$

~~G. Clooney & M. Damon $\xrightarrow{0.33;1}$ Thriller~~

~~G. Clooney $\xrightarrow{0.67;0.8}$ Thriller~~

M. Damon $\xrightarrow{0.33;1}$ Thriller

→ Thriller

Starke Assoziation von Q. Tarantino zu Action (ID 6)!

Tarantino → Action ist ein „small disjunct"

Multi-label Klassifikator (ILAC)

- Modell abhängig von Instanz erzeugen

  ➔ benutze Instanz als Filter

Original-Trainingsdaten:

ID	Label	Title	Actors
1	Drama/Mystery	Solaris	G. Clooney
2	Drama/Thriller	Syriana	G. Clooney, M. Damon
3	Crime/Thriller	Ocean's Eleven	G. Clooney, M. Damon
4	Crime/Thriller	Out of Sight	G. Clooney
5	Thriller	The Peacemaker	G. Clooney
6	Action	Desperado	Q. Tarantino

Testinstanz:

7	?[Action]	From Tusk till Dawn	Q. Tarantino, M. Damon

Gefilterte Trainingsdaten:

ID	Label	Title	Actors
2	Drama/Thriller	Syriana	M. Damon
3	Crime/Thriller	Ocean's Eleven	M. Damon
6	Action	Desperado	Q. Tarantino

Beispiel: ILAC

Gefilterte Trainingsdaten:

ID	Label	Title	Actors
2	Drama/Thriller	Syriana	M. Damon
3	Crime/Thriller	Ocean's Eleven	M. Damon
6	Action	Desperado	Q. Tarantino

Testinstanz:

7	?[Action]	From Tusk till Dawn	Q. Tarantino, M. Damon

s(Thriller)=0,67; s(Action)=0,33

f(Thriller)=1; f(Action)=0,5

→ Thriller/Action

ILAC findet das „small disjunct"

Modell: $\sigma_{min} = 0.2;\ \theta_{min} = 0.67;\ \delta_{min} = 0.5$

$$\text{M. Damon} \xrightarrow{0.67;1} \text{Thriller}$$

$$\text{Q. Tarantino} \xrightarrow{0.33;1} \text{Action}$$

Beispiel: abhängige Label

Testinstanz:

8	**?[Thriller/Crime]**	**Welcome to Collinwood**	**G. Clooney**

Gefilterte Trainingsdaten:

ID	Label	Title	Actors
1	**Drama/Mystery**	**Solaris**	**G. Clooney**
2	**Drama/Thriller**	**Syriana**	**G. Clooney**
3	**Crime/Thriller**	**Ocean's Eleven**	**G. Clooney**
4	**Crime/Thriller**	**Out of Sight**	**G. Clooney**
5	**Thriller**	**The Peacemaker**	**G. Clooney**

Modell: $\sigma_{min} = 0.2;\ \theta_{min} = 0.4;\ \delta_{min} = 0.5$

$$\text{G. Clooney} \xrightarrow{0.8;0.8} \text{Thriller}$$

$$\text{G. Clooney} \xrightarrow{0.4;0.4} \text{Drama}$$

$$\text{G. Clooney} \xrightarrow{0.4;0.4} \text{Crime}$$

s(Thriller)=0,64
s(Drama)=s(Crime)=0,16

f(Thriller)=1; f(Drama)=f(Crime)=0,25 → Thriller

Rang von Drama und Crime gleich bewertet

Abhängigkeiten zwischen Labels nicht aufgedeckt

Multi-label Klassifikator (CLAC)

- Zugewiesene Label als Feature benutzen
- Multi-label class association rule (MCAR)

Regel der Form: $F \cup L \xrightarrow{\sigma,\theta} l_i$

L=Menge von Labels ohne l_i

Progressive label focusing

- Heuristik zur Erforschung des Suchraums für MCARs
  - 1. Iteration:
    - Zu Beginn L=Ø
    - Modell M_1 mit Regeln der Form: $F \xrightarrow{\sigma,\theta} l_i$
    - Bestimme Label l_1
  - 2. Iteration:
    - L={l_1}
    - Modell M_2 mit Regeln der Form: $F \cup \{l_1\} \xrightarrow{\sigma,\theta} l_i$
    - Bestimme Label l_2
  - 3. Iteration:
    - L={l_1, l_2}
    - Modell M_3 mit Regeln der Form: $F \cup \{l_1, l_2\} \xrightarrow{\sigma,\theta} l_i$
    - Bestimme Label l_3
  - … bis keine MCARs mehr gebildet werden können

Beispiel: CLAC

Testinstanz:

8	?[Thriller/Crime]	Welcome to Collinwood	G. Clooney

Gefilterte Trainingsdaten:

ID	Label	Title	Actors
1	Drama/Mystery	Solaris	G. Clooney
2	Drama/Thriller	Syriana	G. Clooney
3	Crime/Thriller	Ocean's Eleven	G. Clooney
4	Crime/Thriller	Out of Sight	G. Clooney
5	Thriller	The Peacemaker	G. Clooney

Modell: $\sigma_{min} = 0.2;\ \theta_{min} = 0.4;\ \delta_{min} = 0.5$

$$\text{G. Clooney} \xrightarrow{0.8;0.8} \text{Thriller}$$

$$\text{G. Clooney} \xrightarrow{0.4;0.4} \text{Drama}$$

$$\text{G. Clooney} \xrightarrow{0.4;0.4} \text{Crime}$$

→ l_1=Thriller

Neue Testinstanz:

8	?[Crime]	Welcome to Collinwood	G. Clooney ^ Thriller

Beispiel: CLAC

Testinstanz:

8	**?[Crime]**	**Welcome to Collinwood**	**G. Clooney ^ Thriller**

Gefilterte Trainingsdaten:

ID	Label	Title	Actors
2	**Drama**	**Syriana**	**G. Clooney ^ Thriller**
3	**Crime**	**Ocean's Eleven**	**G. Clooney ^ Thriller**
4	**Crime**	**Out of Sight**	**G. Clooney ^ Thriller**

Modell: $\sigma_{min} = 0.2;\ \theta_{min} = 0.4;\ \delta_{min} = 0.5$

$$\text{G. Clooney} \wedge \text{Thriller} \xrightarrow{0.67;0.67} \text{Crime}$$

→ l_2=Crime

Keine MCARs mehr → Thriller/Crime

Experimente und Ergebnisse

Datensätze:

Datensatz	Klassifizierungsart	# Instanzen	# Label	Features
ACM-DL (first level)	Textklassifikation	81.251	11	Titel, Abstract, Zitierung, Autoren
ACM-DL (second level)			81	
YEAST	Genklassifikation	2.417	14	Mikroarray-Daten, Phylogenetisches Profil

Evaluationskriterien

- 10 x 10-fold cross-validation
- One-error (O)

  Relative Häufigkeit, dass Instanz dem echten Top-Label nicht zugeordnet wurde
- Hamming loss (H)

  Anteil falsch zugewiesener Labels
- Ranking loss (R)

  Anteil falsch geordneter Labelpaare

Ergebnisse

Ergebnisse für YEAST:

	BoosTexter	ADTBoost.MH	Rank-SVM	IEAC	ILAC	CLAC
O	0.278	0.244	0.217	0.232	**0.213**	**0.213**
H	0.220	0.207	0.196	0.203	0.191	**0.179**
R	0.186	-	0.163	0.178	0.164	**0.150**

Ergebnisse für ACM-DL:

	First Level				Second Level			
	Rank-SVM	IEAC	ILAC	CLAC	Rank-SVM	IEAC	ILAC	CLAC
O	0.244	0.304	**0.238**	**0.238**	0.348	0.427	**0.331**	**0.331**
H	0.225	0.295	0.222	**0.187**	0.327	0.419	0.319	**0.285**
R	0.194	0.276	0.216	**0.179**	0.299	0.378	0.294	**0.273**

Fazit

- Erhöhte Klassifizierungsgenauigkeit durch:
  - Behandlung von „small disjuncts“
  - Untersuchen von Abhängigkeiten zwischen Labels
- Beste Ergebnisse mit CLAC
  - CLAC scheint den bekannten Verfahren überlegen zu sein

Fragen?

Evaluationskriterien

- Ranking-Funktion $f:(\mathrm{X}\times\mathrm{L})\to\mathbf{R}$ bzgl. $f(x,\cdot)$

  → Ranking:

$$R=\{l_1,l_2,...,l_n\} \text{ mit } f(x,l_1)\geq f(x,l_2)\geq ...\geq f(x,l_n)$$

- Testmenge: $T=\langle(x_1,L_1),...,(x_n,L_n)\rangle$

  L_i = echte Labels für Testinstanz x_i

$$\S expr\ddot{} = \begin{cases} 1, \text{ falls expr wahr} \\ 0, \text{ sonst} \end{cases}$$

Evaluationskriterien

- One-error (O)

  Multi-class Klassifikator: $C(x) = arg\,max_l\ f(x,l)$

$$O_T(C) = \frac{1}{n}\sum_{i=1}^{n} \delta\left(C(l_i) \notin L_i\right)$$

- Hamming loss (H)

$$H_T(f) = \frac{1}{|\mathrm{Y}|n}\sum_{i=1}^{n}\sum_{j=1}^{|\mathrm{Y}|}\left(l_j \in f(x_i) \wedge l_j \notin L_i\right) + \left(l_j \notin f(x_i) \wedge l_j \in L_i\right)$$

- Ranking loss (R)

$$R_T(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|L_i||\mathrm{L}-l_i|}\left|\left\{(l_0,l_1) \in L_i \times (\mathrm{L}-L_i) : f(x,l_1) \le f(x,l_0)\right\}\right|$$