Tight Optimistic Estimates for Fast Subgroup Discovery

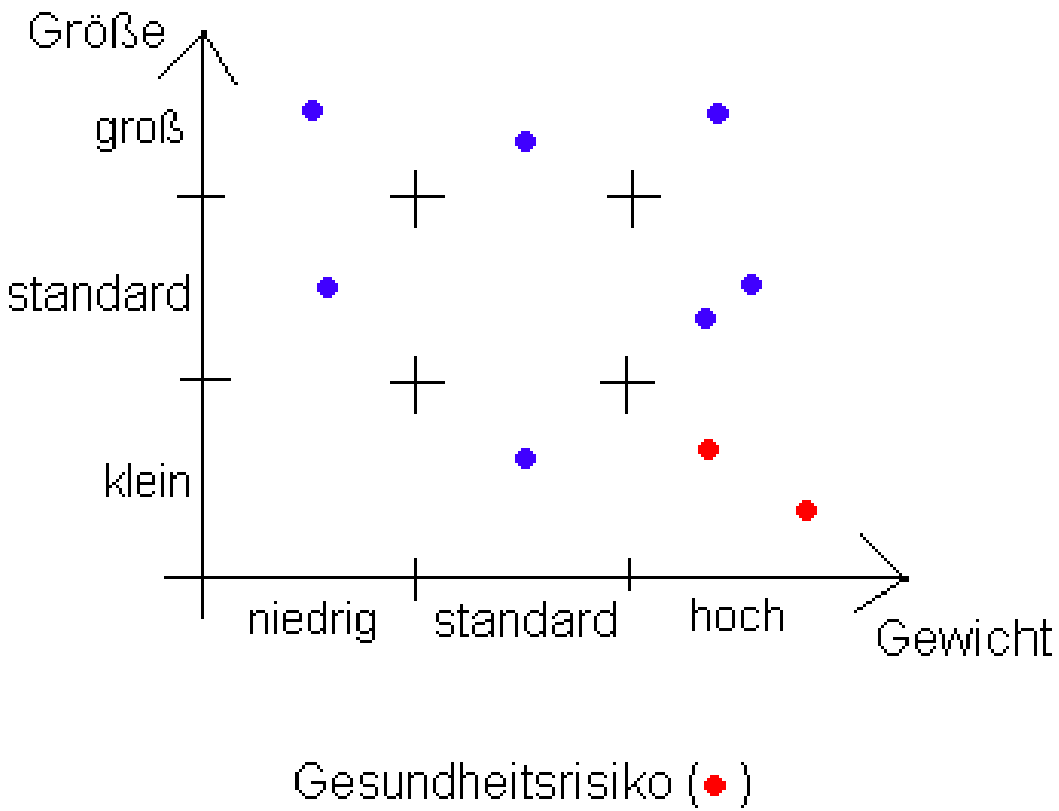
von Henrik Grosskreutz, Stefan Rüping und Stefan Wrobel

Präsentation von Kathrin Rösner

Paper in W. Daelemans et al. (Eds): ECML PKDD 2008, Part I, LNAI 5211, Seiten 440-456, Springer-Verlag

21. Juli 2009

Subgruppen


Größe
groß
standard
klein
niedrig
standard
hoch
Gewicht
Gesundheitsrisiko ( )

Problemstellung

Subgruppen Entdeckung

In einer Population (möglichst große) Subgruppen finden, die unnormal (verteilt) sind

Wieso brauchen wir da besondere Techniken?

- normales Pruning kann nicht benutzt werden

Ansatz von Grosskreutz et al.

- Pruning mit optimistischen Abschätzungen
- Erweiterung auf optimale optimistische Abschätzungen
- Erweiterung auf nicht-binäre Klassen
- Entwicklung einer Familie von optimistischen Abschätzungen die Trade-Off von Rechenzeit und effizienten Abschätzungen erlauben

Inhaltsverzeichnis

Terminologie 1/2

- **Datenbank** $DB = \{R_1, .., R_N\}$
- Reihe R_j ist n-Tupel $(v_{j,1}, ..v_{j,I}, c_j)$
- **Klasse c** $\in \{c_1, ..., c_m\}$
- **Subgruppenbeschreibung sd** $= \{t_1, ..., t_k\}$
- **Term** t_i hat die Form $(a_i = v_i)$
- $sd' = \{t'_1, ..., t'_{k'}\}$ ist eine **Verfeinerung** von $sd = \{t_1, ..., t_k\}$ wenn $\{t'_1, ..., t'_{k'}\} \supseteq \{t_1, ..., t_k\}$
  Symbol: $sd' \succ sd$
- **Subgruppenerweiterung** von sd auf DB:
  alle Reihen aus DB die Terme von sd erfüllen

Terminologie Beispiel

Subgruppenbeschreibung $sd = \{(Groesse = klein)\}$

Gewicht	Größe	Risiko
niedrig	klein	niedrig
niedrig	standard	niedrig
standard	klein	niedrig
standard	groß	niedrig
hoch	klein	hoch
hoch	klein	hoch
hoch	standard	niedrig
hoch	standard	niedrig
hoch	groß	niedrig

Terminologie 2/2

- **Klassenverteilung p** $= (p_1, ..., p_m)$
  mit $p_i := \frac{1}{n} \cdot |\{r | r \in R \wedge class(r) = i\}|$
  und $R = \{R_1, ..., R_n\}$
  Prozentsatz der Reihen mit Klasse i
- **Qualiätsfunktion q**: $DB \times sd \rightarrow \Re$
  Wie interessant ist eine Subgruppe?

Terminologie Beispiel

- Subgruppenbeschreibung $sd = \{(Groesse = klein)\}$
- **Klassenverteilung von sd**
  $p = (p_{niedrig}, p_{hoch}) = (0.5, 0.5)$
- **Klassenverteilung von DB**
  $p_0 = (p_{niedrig}, p_{hoch}) = (\frac{7}{9}, \frac{2}{9})$

Gewicht	Größe	Risiko
niedrig	klein	niedrig
niedrig	standard	niedrig
standard	klein	niedrig
standard	groß	niedrig
hoch	klein	hoch
hoch	klein	hoch
hoch	standard	niedrig
hoch	standard	niedrig
hoch	groß	niedrig

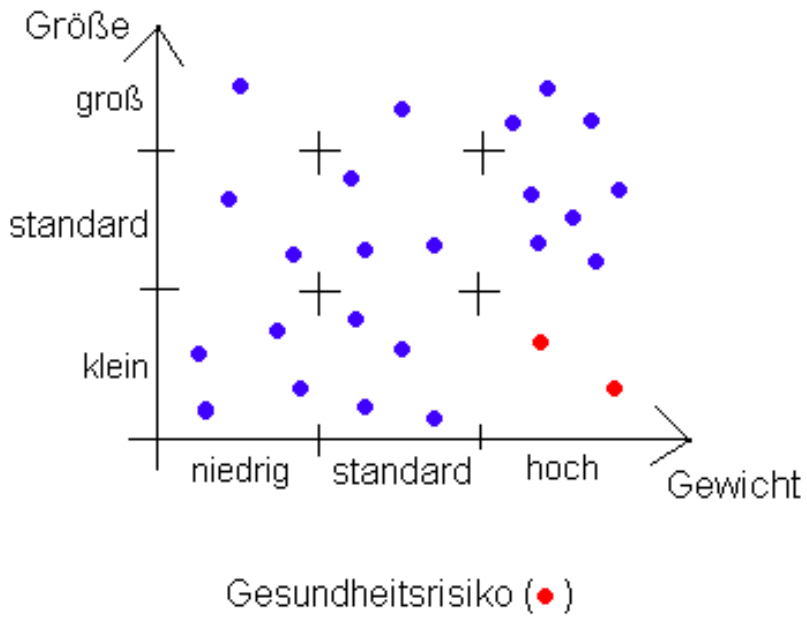
Typische Qualitätsfunktionen für Subgruppen

Name	Type	Definition
Piatetsky-Shapiro	2	$n(p - p_0)$
Split	N	$n \sum_i (p_i - p_{0i})^2$
Gini	N	$\frac{n}{N-n} \sum_i (p_i - p_{0i})^2$
Pearson's χ^2	N	$n \sum_i \frac{(p_i - p_{0i})^2}{p_{0i}}$

Optimistische Abschätzung

Problem beim Pruning des Suchraums: Monotonie gilt nicht für Subgruppen und deren Subgruppen

$sd_1 = (\{Groesse = klein\})$
$sd_2 = (\{Gewicht = hoch\})$
$sd_3 =$
$(\{Groesse = klein, Gewicht = hoch\})$



Optimistische Abschätzung

Definition Optimistische Abschätzung (optimistic estimate)

oe(s) für Qualitätsfunktion q ist eine Funktion, für die gilt:
$\forall$ Subgruppen $s, s' : s' \succ s \Rightarrow oe(s) \geq q(s')$

Alle vorgestellten Qualitätsfunktionen beziehen sich auf Klassenverteilung und Größe der DB ($\mathbf{p}_0$ und $\mathbf{N}$) bzw der Subgruppe ($\mathbf{p}$ und $\mathbf{n}$)

Optimistische Abschätzung

p/n optimistische Abschätzung

ist eine Funktion **oe(** $p(s), n(s), p_0, N$**)**
bezogen auf eine Qualitätsfunktion q, für die gilt:
$\forall Subgruppens, s' : s' \succ s \Rightarrow oe(p(s), n(s), p_0, N) \geq$
$q(p(s'), n(s'), p_0), N)$

Optimistische Abschätzung

Konservative Optimistische Abschätzungen

oe_1 ist konservativer als oe_2 wenn

$\forall N, p_0, n, p : oe_1(p, n, p_0, N) \leq oe_2(p, n, p_0, N)$

Enge (tight) Optimistische Abschätzung

$\forall DB, sd : \exists n', p' :$

$(n' \leq n \wedge n'p' \preceq np \wedge oe*(p, n, p_0, N) = q(p', n', p_0, N))$

- $oe' < oe^* \Rightarrow oe' < oe^* = q(s')$. Widerspruch zu Definition oe ist obere Grenze für $q(s')$
- implizit: oe ist eng wenn es keine echt konservativere Abschätzung gibt
- Merke: es muss keine entsprechende Subgruppenbeschreibung geben, nur eine Untermenge

Enge Optimistische Abschätzung für Piatetsky-Shapiro

Piatetsky-Shapiro Qualitätsfunktion q_{ps}

$$q_{ps} = n(p - p_0)$$

Enge Optimistische Abschätzung für Piatetsky-Shapiro

$$oe^*_{ps} = np(1 - p_0)$$

DB und beliebige Subgruppenerweiterung *s* seien gegeben:

- $p_1 = 0$: immer minimal
- $p_1 > 0$:
  - *s* enthält np_1 Reihen der ersten Klasse
  - Diese Reihen bilden eine Untermenge s' mit Qualität $np_1(1 - p_{0_1})$ (np ist Größe der Untermege, und Verteilung $p_1 = 1$)
  - Es gibt immer so eine Untermenge s'

Enge Optimistische Abschätzung für Piatetsky-Shapiro

Piatetsky-Shapiro Qualitätsfunktion q_{ps}

$$q_{ps} = n(p - p_0)$$

Enge Optimistische Abschätzung für Piatetsky-Shapiro

$$oe^*_{ps} = np(1 - p_0)$$

Gewicht	Größe	Risiko
niedrig	klein	niedrig
niedrig	standard	niedrig
standard	klein	niedrig
standard	groß	niedrig
hoch	klein	hoch
hoch	klein	hoch
hoch	standard	niedrig
hoch	standard	niedrig
hoch	groß	niedrig

Enge Optimistische Abschätzung für Split, Gini und Pearson's χ^2

Kleine Umformulierung:

- $q(m, p_0, N) \equiv q(p, n, p_0, N)$
  - mit $m = (m_1, ..., m_c)$, $c =$ Anzahl Klassen
  - m ist Anzahl der Reihen pro Klasse
  - $m_i = n \cdot (p_1, .., p_c)^T$
  - $n = \sum_j m_j$, $p = \frac{1}{n} \cdot m$

Enge Optimistische Abschätzung für für Multi-Klassen Qualitätsfunktionen Split, Gini und Pearson's χ^2

Enge Optimistische Abschätzung für Multi-Klassen Qualitätsfunktionen Split, Gini und Pearsons's χ^2

$oe_q^*(p_1, ..., p_c, n, p_0, N) :=$
$\max_{m'_1,...,m'_c | m'_i \in \{0, np_i\}} \left\{ q((m'_1, ..., m'_c)^T, p_0, N) \right\}$

- Enge OA für beliebige Qualitätsfunktion q:
  $\max_{m'_1,...,m'_c | \forall i: m'_i \in \mathbf{N}_+ \wedge 0 \leq m'_i \leq np_i} \left\{ q((m'_1, ..., m'_c)^T, p_0, N) \right\}$
- m ist Anzahl der Reihen pro Klasse
- zZ: beide Funktionen sind Äquivalent

Enge Optimistische Abschätzung für für Multi-Klassen Qualitätsfunktionen Split, Gini und Pearson's χ^2

Enge Optimistische Abschätzung für Multi-Klassen Qualitätsfunktionen Split, Gini und Pearsons's χ^2

$oe_q^*(p_1, ..., p_c, n, p_0, N) :=$
$\max_{m'_1,...,m'_c | m'_i \in \{0, np_i\}} \left\{ q((m'_1, ..., m'_c)^T, p_0, N) \right\}$

- Enge OA für beliebige Qualitätsfunktion q:
  $\max_{m'_1,...,m'_c | \forall i: m'_i \in \mathbf{N}_+ \wedge 0 \leq m'_i \leq np_i} \left\{ q((m'_1, ..., m'_c)^T, p_0, N) \right\}$
- Jede Qualitätsfunktion q aus dieser Präsentation ist konvex über m
- Für jedes q ist das Maximum über ein Polyhedron $P = [0, m_1] \times ... \times [0, m_c]$ ein Extrempunkt von P

Enge Optimistische Abschätzung für für Multi-Klassen Qualitätsfunktionen Split, Gini und Pearson's χ^2

Enge Optimistische Abschätzung für Multi-Klassen Qualitätsfunktionen Split, Gini und Pearsons's χ^2

$oe_q^*(p_1, ..., p_c, n, p_0, N) :=$
$\max_{m'_1,...,m'_c | m'_i \in \{0, np_i\}} \left\{ q((m'_1, ..., m'_c)^T, p_0, N) \right\}$

- Laufzeit von $O(c2^c)$
- Jede Qualitätsfunktion aus dieser Präsentation kann umgeschrieben werden zu
  $q = \phi_1 + ... + \phi_c$
- Unabhängig von Größe der Datenbank

Familie von ansteigend konservativen optimistischen Abschätzungen

Definition oe_p^d

$$oe_p^d(p, n, p_0, N) := \sum_{j=1,d+1,2d+1,\ldots} (\max_{m'_j,\ldots,m'_{j+d-1} | m'_j \in \{0, np_j\}} (max \left\{ \sum_{i=j}^{j+d-1} \phi_i(m'_-, p_0, N), \sum_{i=j}^{j+d-1} \phi_i(m'_+, p_0, N) \right\}))$$

- Idee: Nur $d < c$ verschiedene Klassen genau ausrechnen, für restliche Klassen obere Schranken nutzen
- Laufzeit: $O(c2^d)$ statt $O(c2^c)$
- m'_- heisst $m_j = 0$ fuer alle j die nicht in den d betrachteten Klassen sind, m'_+ analog

Beweis: oe_p^d ist obere Grenze für Qualitätsfunktionen

$$\max_{m'_1,\ldots,m'_c|\forall i.m'_i\in N_+\wedge 0\leq m'_i\leq np_i}\{\sum_{i=1}^{c}\phi_i(\boldsymbol{m'},\boldsymbol{p_0},N)\}=$$

$$\max_{m'_1,\ldots,m'_c|\forall i.m'_i\in\{0,np_i\}}\{\sum_{i=1}^{c}\phi_i(\boldsymbol{m'},\boldsymbol{p_0},N)\}\leq$$

$$\sum_{j=1,d+1,2d+1,\ldots,c}\max_{m'_1,\ldots,m'_c|\forall i.m'_i\in\{0,np_i\}}\{\sum_{i=j}^{j+d-1}\phi_i(\boldsymbol{m'},\boldsymbol{p_0},N)\}\leq$$

$$\sum_{j=1,d+1,\ldots,c}\max_{m'_j,\ldots,m'_{j+d-1}|m'_j\in\{0,np_j\}}\left[\max_{m'_1,\ldots m'_{j-1},m_{j+d},\ldots m'_c|m'_i\in\{0,np_i\}}\{\sum_{i=j}^{j+d-1}\ldots\}\right]\leq$$

$$\sum_{j=1,d+1,\ldots,c}\left[\max_{m'_j,\ldots,m'_{j+d-1}|m'_j\in\{0,np_j\}}\left(\max_{(\forall k.\,m'_k=0),(\forall k.\,m'_k=np_k)}\sum_{i=j}^{j+d-1}\phi_i(\boldsymbol{m'},\boldsymbol{p_0},N)\right)\right]$$

□

Beweis: Eigenschaften von oe_p^d

oe_p^d ist eng wenn $d \geq c$

Alle c Klassen werden in oe_p^d betrachtet

oe_p^d ist eng wenn $c = 2$ und $d \geq 1$

oe_p^d arbeitet mit 2 Indexmengen:
$\{0, ..., j-1, j+d, ..., c\}$ und $\{j, ..., j+d-1\}$
Alle Kombinationen der beiden Klassen werden betrachtet.

oe_p^{2d} ist mindestens so konservativ wie oe_p^d

oe_p^{2d} betrachtet mind. alle Klassen-Kombinationen, die oe_p^d betrachtet

Beweis: Eigenschaften von oe_p^d

oe_p^d ist genau dann eng, wenn ($d \geq c$ oder ($c = 2$ und $d \geq 1$)

- Beweis für Enge bei Fällen ($d \geq c$) und ($c = 2$ und $d \geq 1$) bereits erbracht
- Beweis dafür, dass oe_p^d sonst nicht eng ist:
  - $p_0 = (0.1, 0.45, 0.45)$
    bzw. $p_0 = (0.1, \frac{0.3}{c-3}, ..., \frac{0.3}{c-3}, 0.3)$ für $c > 3$
  - Subgruppe s habe $m = (10, 10, 0, ..., 0, 10)$
  - Untermenge m' mit höchster Qualität ist bei
    $m' = (10, 0, 0, ..., 0, 0)$ (höchste Verschiedenheit)
  - Der letzte Summand für oe_p^d würde $m'_c = 10 \vee m'_c = 0$ nur mit $m'_1 = m'_2$ betrachten

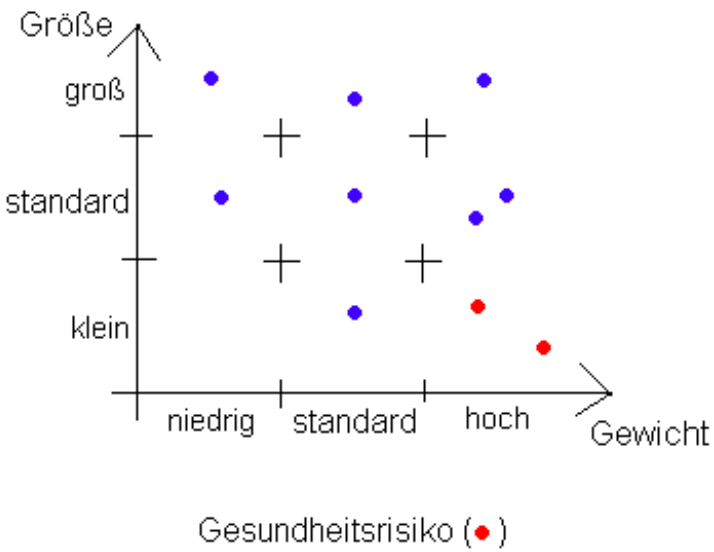
Algorithmus

Algorithmus 'DpSubgroup'

- Branch-and-Bound DFS Algorithmus
- Optimistische Abschätzungen zum Verkleinern des Suchraumes (Pruning)
- Optimistische Abschätzungen als Entscheidungshilfe für Reihenfolge der Knoten im DFS
- Implementation mit FP-Bäumen (verschnellert Berechnung von p und n einer Subgruppe)

Beispielrechnung oe*

$$oe^*_{split}(p_1, \ldots, p_c, n, p_0, N) :=$$
$$\max_{m'_1,\ldots,m'_c | m'_i \in \{0, np_i\}} \left\{ q_{split}((m'_1, \ldots, m'_c)^T, p_0, N) \right\}$$



Gewicht niedrig	(0,0), (2,0)
Gewicht standard	(0,0), (3,0)
Gewicht hoch	(0,0), (3,0), (0,2), (3,2)
Größe klein	(0,0), (1,0), (0,2), (1,2)
Größe standard	(0,0), (4,0)
Größe groß	(0,0), (3,0)

Beispielrechnung oe*

$$q_{split} = n \sum_i (p_i - p_{0_i})^2$$

(m1,m2)	q
(0,0)	0
(1,0)	0.08
(2,0)	0.16
(3,0)	0.24
(4,0)	0.32
(0,2)	2.56
(1,2)	0.73
(3,2)	0.4

$$oe^*_{split}(p_1, ..., p_c, n, p_0, N) := \max_{m'_1,...,m'_c | m'_i \in \{0, np_i\}} \{q_{split}(...)\}$$

	q	oe*
Gewicht niedrig	0.16	0.16
Gewicht standard	0.24	0.24
Gewicht hoch	0.4	2.56
Größe klein	1.3	2.56
Größe standard	0.32	0.32
Größe groß	0.24	0.24

Algorithmus

Procedure DpSubgroup
Input: FPTree F, Subgroup s^*, Double $minQ$
OE:= new associative Array (InitializedWith $-\infty$)
SG:= new associative Array (used to store refinements)
for all t in F.`getTerms`() **do**
  s := `createRefinement`(s^*, t)
  SG[t] := s
  if `getG`(s, F) $\geq$ $minG$ **then**
    if `getQ`(s, F) $\geq$ $minQ$ **then**
      `addToResultQueue`(s)
      if *resultQueue* contains k subgroups **then**
        $minQ$:= Max($minQ$, `getWorstResQlty`())
      end if
    end if
    $OE[t]$ = `optEstim`(s, F)
  end if
end for
if Depth(s^*)+1<$maxDepth$ **then**
  $tOrdered$:= `getTermsInRecurseOrder`(F, ...)
  for all t in $tOrdered$ **do**
    if $OE[t] \geq minQ$ **then**
      ct = `getCondTree`(t, F, OE, $minQ$)
      $minQ$= DpSubgroup(ct, $SG[t]$, $minQ$)
    end if
  end for
end if
Output: $minQ$

Experimentelle Ergebnisse

	oe_q^2	oe_q^3	oe_q^4	oe_q^*
Minimal relative runtime compared to oe_q^1	62%	21%	3%	1%
Average relative runtime compared to oe_q^1	93%	69%	63%	847%
Maximal relative runtime compared to oe_q^1	113%	118%	135%	3640%

Quellen

- H. Grosskreutz, S. Rüping, S. Wrobel:
  Tight Optimistic Estimates for Fast Subgroup Discovery.
  In W. Daelemans et al. (Eds): ECML PKDD 2008, Part I,
  LNAI 5211, Seiten 440-456, Springer-Verlag
- Grosskreutz, H., Rüping, S., Shaabani, N., Wrobel, S.:
  Optimistic estimate pruning strategies for fast exhaustive
  subgroup discovery.
  Technical report, Fraunhofer Institute IAIS (2008),
  http://publica.
  fraunhofer.de/eprints/urn:nbn:de:0011-n-723406.pdf