Vorlesung Wissensentdeckung

Knowledge based sampling – KBS

Katharina Morik, Claus Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

30.06.2015

Gliederung

Lernaufgabe Subgruppenentdeckung

- Gegeben
  - X der Raum möglicher Beobachtungen mit einer Wahrscheinlichkeitsverteilung D,
  - $S \subseteq X$ eine gemäß D gezogene Stichprobe,
  - L_H der Raum möglicherweise gültiger Regeln, wobei jeder Regel $h \in L_H$ eine Extension zugeordnet ist: $ext(h) \subseteq X$ und
  - eine Qualitätsfunktion

$$q : L_H \rightarrow \mathcal{R}$$

- finde
  - eine Menge $H \subseteq L_H, \mid H \mid = k$
  - und es gibt keine $h' \in L_H \backslash H, h \in H$, für die gilt $q(h') \geq q(h)$

Beispiel der Subgruppenentdeckung

Es werden Gruppen beschrieben, die sich abweichend von der Gesamtpopulation verhalten.
Es geht nicht notwendigerweise um Vorhersage, sondern um Beschreibung! Trotzdem ist meist eine Hypothese eine Abbildung $h : X \rightarrow Y$.

- Unter den alleinstehenden jungen Männern in ländlichen Regionen ist der Anteil an Lebensversicherungskinden signifikant niedriger als im gesamten Kundenbestand.
- Verheiratete Männer mit Pkws der Luxusklasse machen nur 2 Prozent der Kunden aus, erzeugen aber 14 Prozent der Lebensversicherungsabschlusssumme.

Ansätze zur Subgruppenentdeckung

- Aufzählend: vollständige Suche im strukturierten Raum L_H mit Pruning – Garantie, dass die k besten Regeln gefunden werden.
  Explora (Klösgen 1996), Midos (Wrobel 1997)
- Heuristisch: ein Entscheidungsbaumlerner wird so verändert, dass seine Qualitätsfunktion die der Subgruppenentdeckung wird und Beispiele ein veränderliches Gewicht erhalten – keinerlei Garantie.
  CN2-SD (Lavrac et al. 2004)
- Probabilistisch: Stichproben-bezogene Fehler werden während das Scans der Daten abgeschätzt – probabilistische Garantie, dass die k besten Regeln gefunden werden.
  (Scheffer, Wrobel 2002)

Modellselektion

- Die Menge H der gewählten Hypothesen kann auch als Modell betrachtet werden.
- Die Subgruppenentdeckung ist dann ein Problem der Modellselektion.
- Dabei geht es immer um Gütekriterien.
- Wir hatten ja schon:
  - Accuracy
  - Precision
  - Recall
  - Mittlerer quadratischer Fehler, quadratische Fehlersumme, erwarteter quadratischer Fehler, 0-1-Verlust
  - Maximum Likelihood
  - Entropie
  - Bayes Information Criterion
  - Minimum Description Length

Lift

- Für eine Regel $h = A \rightarrow Y$, wobei A eine Menge von Literalen ist und $Y = \{0, 1\}$ ist

$$Lift(A \rightarrow Y) = \frac{Pr[A, Y]}{Pr[A] \cdot Pr[Y]} = \frac{precision(A \rightarrow Y)}{Pr[Y]} \quad (1)$$

- Bei $Lift(A \rightarrow Y) = 1$ sind A und Y unabhängig.
- Bei $Lift(A \rightarrow Y) > 1$ steigt die bedingte Wahrscheinlichkeit für Y gegeben A.
- Bei $Lift(A \rightarrow Y) < 1$ sinkt die bedingte Wahrscheinlichkeit für Y gegeben A.
- Lift normalisiert die precision gegenüber einer verzerrten Verteilung der Klassen!

Coverage und Bias von Regeln

- Die Wahrscheinlichkeit, dass der Antezedens A der Regel auf ein Beispiel zutrifft bei einer Verteilung D der Beispiele ist:

$$Cov(A \rightarrow Y) = Pr[A]$$

- Die Differenz zwischen der bedingten Wahrscheinlichkeit von Y gegeben A und der a priori Wahrscheinlichkeit für Y ist der Bias:

$$Bias(A \rightarrow Y) = Pr[Y \mid A] - Pr[Y] = Pr[Y] \cdot (Lift(A \rightarrow Y) - 1)$$

Weighted relative accuracy WRAcc

- Man kann Bias und Coverage für eine Anwendung mit einem Parameter α geeignet gewichten.
  - Vielleicht will man auf jeden Fall alles abdecken, weil man alle Beispiele irgendwie behandeln muss. Dann gewichtet man Coverage hoch.
  - Vielleicht findet man nur Abweichungen von der a priori Wahrscheinlichkeit interessant. Dann gewichtet man Bias hoch.
  - Bei gleichgewichteten Coverage und Bias $\alpha = 0,5$ erhält man das selbe Ergebnis wie beim binominalen Test, der die Nullhypothese (A hat keinen Einfluss) testet.
- Für eine Regel h und eine Gewichtung $\alpha \in [0, 1]$ ist

$$WRAcc(\alpha, h) = Cov(h) \cdot Bias(h)$$

Wofür die Maße?

- Jetzt wissen wir, wie wir Regeln auswählen können.
- Wir wollen aber auch noch wissen, wie gut das Modell, also die gesamten Regeln, ist.
- Dann können wir die Regelmenge auswählen, die am besten ist.

Sensitivität und Spezifität – ROC

- Sensitivität (Recall): Wahrscheinlichkeit, dass ein positives Beispiel auch als positiv erkannt wird. (TP: true positives)
- Spezifizität: Wahrscheinlichkeit, dass ein negatives Beispiel auch als negativ erkannt wird. (TN: true negatives)
- Die Receiver Operator Characteristic (ROC) Kurve setzt Sensitivität und Spezifizität in Beziehung für verschiedene Parameter. Je nach Benutzerinteresse (TP wichtiger? TF wichtiger? Beides?) wird das Modell gewählt.

Beispiel

Tabelle: $y = 1$ für Spam, Fehler insgesamt 9%

	Predicted	
True	email	spam
email	57,3	4,0
spam	5,3	33,4

Sensitivität:

$$100 \cdot \frac{33,4}{33,4 + 5,3} = 86,3\%$$

Spezifizität:

$$100 \cdot \frac{57,3}{57,3 + 4,0} = 93,4\%$$

ROC im Bild

Ein Parameter wurde zwischen 0,1 und 10 variiert.

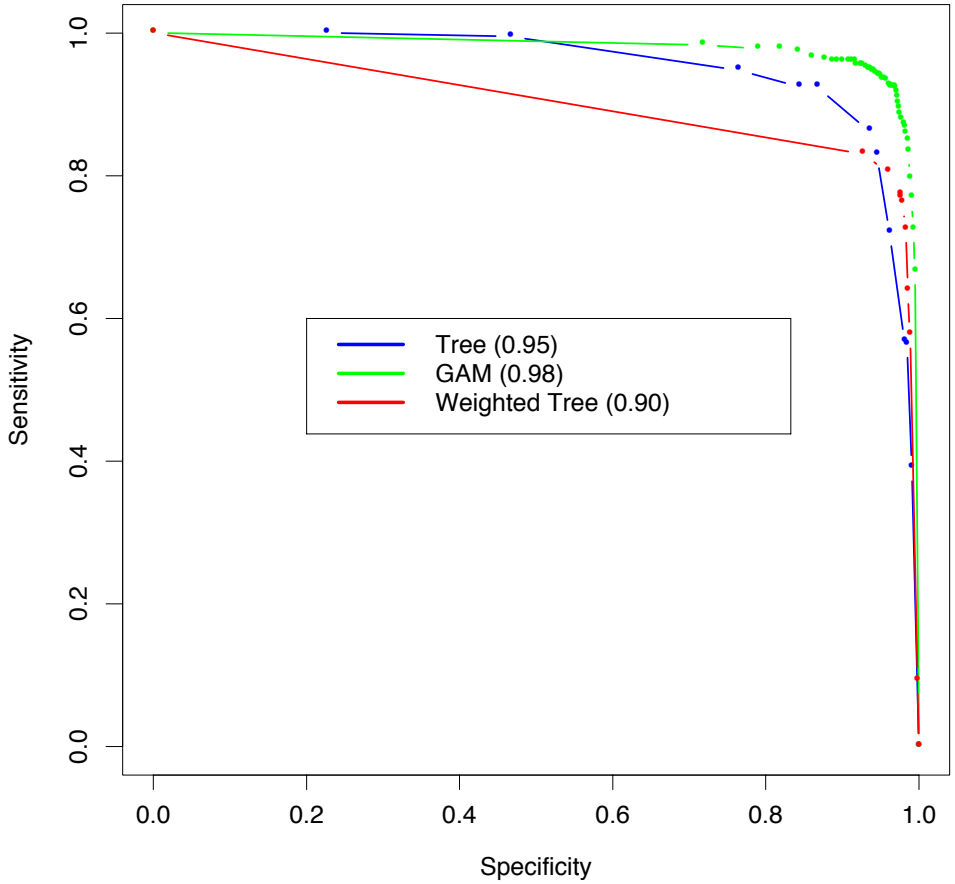

Figure 9.6: *ROC curves for the classification rules fit to the* `spam` *data. Curves that are closer to the north-east corner represent better classifiers. In this case the*

Area Under the Curve

AUC

Für $h : X \to Y, Y \in \{0, 1\}$ und $D : X \times Y \to \mathcal{R}^+$ ist die Area Under the ROC Curve (AUC) die Wahrscheinlichkeit

$$AUC(h) = Pr[h(\vec{x}) \geq_q h(\vec{x}') \mid y = 1, y' = 0]$$

dass ein zufällig gezogenes positives Beispiel höher bewertet wird gemäß einer Qualitätsfunktion q als ein zufällig gezogenes negatives Beispiel.
AUC ist invariant gegenüber monotonen Transformationen von h.

Idee eines Algorithmus', der AUC berechnet

- Schätze für jedes Beispiel in S die Wahrscheinlichkeit, ein positives zu sein.
- Ordne die Beispiele nach dieser Wahrscheinlichkeit (ranking).
- Bewerte ein Lernergebnis nach der Anzahl $\Lambda(h, S)$ der notwendigen Vertauschungen der Reihenfolge (des rankings).
- Sei S_+ die Menge der positiven Beispiele, S_- die Menge der negativen, dann ist

$$AUC(h, S) = \frac{\Lambda(h, S)}{\mid S_+ \mid \cdot \mid S_- \mid}$$

Abhängigkeit des Lernergebnisses von S

- Eigentlich wollen wir ja ein optimales (oder wenigstens angenähert optimales) Lernergebnis auch für noch nicht gesehene Beispiele haben.
- Die ROC Kurve bezieht sich wie auch AUC nur auf die Stichprobe S.
- Meist sind die Datenmengen so groß, dass wir nur eine Stichprobe behandeln können.
- Wir wollen jetzt eine Stichprobe ziehen, die ungefähr so verteilt ist wie die Gesamtmenge.
- Leider haben wir keine Ahnung, was die wahre Verteilung ist!

i.i.d. erhaltende Stichprobe

- Die Daten insgesamt, $\mathbf{X}$, und die Stichprobe S sollen i.i.d. verteilt sein.
- Folgen von Zufallsvariablen, die sowohl unabhängig als auch identisch verteilt sind werden üblicherweise mit i.i.d. (für independent and identically distributed) bezeichnet.
  - Beispiel dreimaliges Würfeln:
    - X_1 1. Wurf, X_2 2. Wurf, X_3 3. Wurf sind i.i.d. verteilt.
    - $X_4 = X_1 + X_2$ und $X_5 = X_2 + X_3$ sind zwar identisch verteilt, aber nicht unabhängig.
    - X_4 und X_3 sind unabhängig, aber nicht identisch verteilt.
- Wenn die Daten in der Datenbank in zufälliger Reihenfolge gespeichert sind, ergibt das Ziehen der m ersten Daten eine i.i.d. erhaltende Stichprobe.

Ziehen der Stichprobe mit/ohne Zurücklegen

- Zufällig ein Beispiel ziehen ist Ziehen mit Zurücklegen. Dabei kann es Doppelte geben und damit eine Verzerrug (Bias). Die Wahrscheinlichkeit für Doppelte beim Ziehen von m Beispielen aus N ist:

$$p_m = \frac{N!}{(N-m)! \cdot N^m}$$

  Also sinkt die Wahrscheinlichkeit, keine Doppelten zu haben, $1 - p_m$, exponentiell mit Steigen von m.
- Zufällig ein Beispiel ziehen und es nicht Zurücklegen verfälscht nicht: jedes Beispiel hat die selbe Wahrscheinlichkeit, gezogen zu werden m/N. Leider ist dies aufwändig: man muss prüfen, ob ein Beispiel der Datenbank schon gezogen wurde, logarithmische Laufzeit.

Konfidenz

- Wir möchten gern wissen, bei wie vielen Beispielen wir wie sicher sein können, uns nicht zu verschätzen.
- Dazu nehmen wir einen Konfidenzwert δ und Schranken für die Wahrscheinlichkeit.
- Dann können wir nach und nach imer größere Stichproben ziehen, bis wir uns sicher genug sind. Und dann aufhören!

Chernoff-Schranke

- Sei p die Wahrscheinlichkeit, dass ein Beispiel gezogen wird, das von einer Regel h korrekt klassifiziert wird.
- Bei i.i.d. Stichproben ist p konstant für alle Regeln.
- Die Zufallsvariable X_i, mit $i = 1, ..., m$ sei 1 für die korrekte Klassifikation, 0 sonst.
- Der Erwartungswert für $\bar{Y} = 1/m \sum X_i$ ist gerade p: $E(\bar{X}) = p$
- Die Standardabweichung ist $\sigma(\bar{Y}) = \sqrt{\frac{p(1-p)}{m}}$
- Die Chernoff-Schranke sagt für beliebigen Parameter λ:

$$Pr[\bar{Y} \geq (1 + \lambda)p] \leq exp(-\lambda^2 mp/3) \quad (2)$$

$$Pr[\bar{Y} \leq (1 - \lambda)p] \leq exp(-\lambda^2 mp/2) \quad (3)$$

Chernoff-Schranke zur Abschätzung der geeigneten Stichprobengröße – Beispiel

- Wie wahrscheinlich ist es, dass Regeln mit der wahren Accuracy $Acc = p = 75\%$ bei einer Stichprobe der Größe m nicht besser als reiner Zufall abschneiden?
- Sei $\bar{Y} = \widehat{Acc}$ der Anteil korrekter Klassifikationen und der reine Zufall 50%. $\lambda = 1/3$, weil $(1 - 1/3) \cdot Acc = 50\%$.
- Wegen Gleichung (2) ergibt sich:

$$Pr[\widehat{Acc} \leq (1 - 1/3) \cdot Acc] \leq exp(-(1/3)^2 m \cdot Acc/2)$$
$$\Leftrightarrow Pr[\widehat{Acc} \leq 1/2] \leq exp(-1/9m3/8) = exp(-\frac{m}{24})$$

- Risiko $\leq \delta = 5\%$, dass bei $m \geq 72$ Beispielen ein 75% gutes h die Hälfte falsch klassifiziert:

$$exp(-\frac{m}{24}) \leq \delta \Leftrightarrow -\frac{m}{24} \leq ln\,\delta = -ln\frac{1}{\delta} \Leftrightarrow m \geq 24ln\frac{1}{\delta} = 24ln20$$

Hoeffding-Schranke

- Die Hoeffding-Schranke ist unabhängig von Acc definiert.

$$\begin{aligned} Pr[\bar{Y} - p \geq \epsilon] &\leq exp(-2\epsilon^2 m) \\ Pr[\bar{Y} - p \leq -\epsilon] &\leq exp(-2\epsilon^2 m) \\ Pr[\mid \bar{Y} - -p \mid \geq \epsilon] &\leq 2exp(-2\epsilon^2 m) \end{aligned} \quad (4)$$

- Die wahre Acc soll um nicht mehr als 10% über- oder unterschätzt werden. Wegen Gleichung (4) ergibt sich:

$$Pr[\mid \widehat{Acc} - Acc \mid \geq 0,1] \leq 2exp(-2 \cdot (0,1)^2 m) \leq 2exp(0,02m)$$

- Risiko $\leq \delta = 5\%$ dafür bei $m \sim 184$ Beispielen:

$$2exp(-0,02m) \leq 0,05 \Leftrightarrow -0,02m \leq ln\frac{1}{40}$$
$$\Leftrightarrow 0,02m \geq ln\ 40 \Leftrightarrow m \geq 50ln\ 40 \sim 184$$

Stichprobengröße für Subgruppenentdeckung

- Sei Konfidenzparameter $\delta \in [0, 1]$ und höchster geduldeter Fehler $\epsilon \in \mathcal{R}^+$, es sollen die k besten Regeln $H \in L_H$ gemäß einer Qualitätsfunktion q so gelernt werden, dass mit einer Wahrscheinlichkeit von mindestens $1 - \delta$ eine i.i.d. Stichprobe $\mid S \mid = m$ die wahre Qualität $\hat{q}$ höchstens um $E(m, \delta)$ verfälscht.
- In (Scheffer, Wrobel 2002) wird für die verschiedenen Qualitätskriterien aufgeführt, was $E(m, \delta)$ ist.
- Für Acc kann man die worst case Größenordnung der Stichprobe durch die Menge betrachteter Regeln L_H angeben:

$$m = O(\frac{1}{\epsilon^2} log \frac{\mid L_H \mid}{\delta})$$

Generic Sequential Sampling (Scheffer, Wrobel 2002)

- Durchgehen der Beispiele (scan) bis höchstens $m = O(\frac{1}{\epsilon^2} log \frac{|L_H|}{\delta})$ betrachtet wurden;
  1. Cov für positive und negative Beispiele bestimmen;
  2. Anordnen der Regeln nach dem Qualitätskriterium (ranking);
  3. Alle Regeln aussortieren aus dem Lernergebnis H, wenn sie häufiger als $\delta(2m \mid L_H \mid)$ falsch waren; die Wahrscheinlichkeit, eine gute Regel auszusortieren, ist dann höchstens $\delta/2$.
  4. Wenn $\mid H \mid \leq k$, wird H ausgegeben und die Regeln sind mit einer Wahrscheinlichkeit von mindestens $1 - \delta$ bis auf eine Abweichung von höchstens ϵ optimal.

Stratifizierte Stichproben

Stratifizierte Dichtefunktion

Für $D : X \times Y \rightarrow \mathcal{R}^+$ ist die stratifizierte Dichtefunktion D' definiert als

$$D'(x, y) = \frac{D(x, y)}{\mid Y \mid \cdot Pr[y = y']}$$

und falls wir klassifizieren mit $f : X \rightarrow Y$ als

$$D'(x, y) = \frac{D(x)}{\mid Y \mid \cdot Pr[f(x)]}$$

.

Es wird also die gegebene Verteilung D so geändert, dass die Verteilung der Klassen in D' gleich ist.

Ergebnis von Scholz 2005

- Wenn stratifizierte Stichproben gezogen, d.h. die Verteilung ensprechend geändert wird, entspricht die Subgruppenentdeckung mit der Qualitätsfunktion $WRAcc$ genau einer Klassifikation mit der Gütefunktion Acc.
- Man kann also in Ruhe die Lernalgorithmen für Klassifikation verwenden und braucht keine neuen zu erfinden.
- Allerdings muss man eine Stratifizierung, also Veränderung der Verteilung algorithmisch formulieren.
- Idee: Das tut man beim Ziehen von Stichproben.
- Folge: das Lernen auch aus großen Datenmengen geht schnell!

Knowledge-Based Sampling for Subgroup Discovery

- Wir wollen Vorwissen berücksichtigen, insbesondere nicht redundante Regelmengen H lernen. Dabei ist die Redundanz der Extension wichtig, nicht, dass sie durch verschiedene Merkmale ausgedrückt werden.
- Auch bereits gelernte Regeln $h \in H$ sind Vorwissen.
- Wir wollen wenig Beispiele bearbeiten müssen.
- Wir wollen vorhandene Algorithmen nutzen.
- Wir wollen diejenigen Subgruppen zurückliefern, die von der Allgemeinheit abweichen.
- Meist interessiert den Anwender die Extension einer solchen abweichenden Gruppe.

Martin Scholz *Scalable and Accurate Knowledge Discovery in Real-World Databases*, Dissertation am LS8, TU Dortmund, 2006

Ansatz: die Verteilung verändern

Die neue Verteilung D' soll nichts Wesentliches verändern:

$$Pr_{D'}[x \mid A, Y] = Pr_D[x \mid A, Y] \quad (5)$$

$$Pr_{D'}[x \mid A, \neg Y] = Pr_D[x \mid A, \neg Y] \quad (6)$$

$$Pr_{D'}[x \mid \neg A, Y] = [x \mid \neg A, Y] \quad (7)$$

$$Pr_{D'}[x \mid \neg A, \neg Y] = [x \mid \neg A, \neg Y] \quad (8)$$

Die Beschränkungen (5 – 8) bestimmen die neue Verteilung $D' : X \to \mathcal{R}^+$ eindeutig:

$$Pr_{D'}(x) = Pr_D(x) \cdot (Lift_D(h, x))^{-1} \quad (9)$$

$Lift(h, x)$

Der Lift eines Beispiels $x \in X$ ist für eine Regel $A \to Y$:

$$Lift(A \to Y, x) = \begin{cases} Lift(A \to Y), falls & x \in ext(A) \cap ext(Y) \\ Lift(A \to \neg Y), falls & x \in ext(A) \cap ext(\neg Y) \\ Lift(\neg A \to Y), falls & x \in ext(\neg A) \cap ext(Y) \\ Lift(\neg A \to \neg Y), falls & x \in ext(\neg A) \cap ext(\neg Y) \end{cases}$$

Lift drückt genau aus, wie weit eine Gruppe A von der allgemeinen Verteilung von Y abweicht.

Knowledge-Based Sampling für Subgruppenentdeckung

Gegeben $X = \{(x_1, y_1), ..., (x_N, y_N)\}$ und k, finde eine Menge $H = \{h_1, ..., h_k\}$

1. Stelle die a priori Verteilung $\pi(y)$ für jedes $y \in Y$ fest.
2. Stratifizieren der Verteilung: $D_1(x_i) = \pi(y_i)^{-1}$ für $i = 1, ..., N$
3. für $t = 1$ bis k do
   - $h_t = RegelLernen(D_t, X)$
   - Kontingenzmatrix für h_t mit Gewichten gemäß D_t
   - Lift-Bewertung für h_t gemäß der Kontingenzmatrix
   - $D_{t+1}(x_i) = D_t(x_i) \cdot (Lift_{Dt}(h_t, x))^{-1}$ für $i \in \{1, ..., N\}$
4. Ausgabe $\{h_1, ..., h_k\}$ mit $Lift(h_i)$ (Definition 1)

Subgruppen für die Vorhersage

- Die Regeln können mit ihrer Gewichtung zu einem Ensemble zusammengefasst werden.
- LiftRatio LR:

$$LR(A \to Y, x) = \begin{cases} \frac{Lift(A \to Y)}{Lift(A \to \neg Y)}, falls & x \in ext(A) \\ \frac{Lift(\neg A \to Y)}{Lift(\neg A \to \neg Y)}, falls & x \in ext(\neg A) \end{cases} \quad (11)$$

- Für alle Regeln, wobei D_0 die uniforme Verteilung über X ist:

$$\hat{\beta}(x) = \frac{Pr_{D0}[Y]}{Pr_{D0}[\neg Y]} \cdot \prod_{1 \leq i \leq k} LR_{Di}[(A^i \to Y), x] \quad (12)$$

Was wissen Sie jetzt?

- Sie haben eine neue Lernaufgabe kennengelernt: Subgruppenentdeckung.
- Wie bisher bei (fast) jeder Lernaufgabe, ging es gleich um Modellselektion. Hier für eine Menge von Hypothesen (Regeln), nicht eine Funktion.
- Sie haben neue Gütekriterien kennengelernt: Lift, WRAcc, Spezifizität und Sensitivität
- Für eine Reihe von Experimenten haben Sie ROC und AUC kennengelernt.
- Die Größe von Stichproben in Bezug auf das Risiko, dass das Lernergebnis falsch ist, wurde mt Chernoff und Hoeffding beschränkt.

Und Sie wissen noch mehr!

- Zwei effiziente Ansätze zur Subgruppenentdeckung, von Wrobel und von Scholz, beruhen darauf, dass man nicht alle Beispiele zu betrachten braucht.
- Sie kennen Knowledge-Based Sampling für Subgruppenentdeckung und wie man das Ergebnis für die Klassifikation verwenden kann.