Data Mining und Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

5. Übungsblatt

Aufgabe 1 Covering-Algorithmus und Coverage-Space

Visualisieren Sie den Ablauf des Covering-Algorithmus mit den Daten des letzten Übungsblatts Aufgabe 1b). Veranschaulichen Sie das Lernen jeder einzelnen Regel im Coverage-Space. Zeichnen Sie auch alle untersuchten Kandidaten-Regeln ein und skizzieren Sie zusätzlich die Linien, die dem jeweiligen Bewertungsmaß entsprechen. Sie sollten sowohl einen Graphen für jede Regel als auch für das Lernen der gesamten Theorie anfertigen

- für das Bewertungsmaß Accuracy, wobei die Regel mit der höchsten Bewertung ausgewählt wird und
- für das Bewertungsmaß Precision (zumindest für die zweite gelernte Regel, da die erste Regel nur einmal verfeinert wird).

Aufgabe 2 Heuristiken und Äquivalenzen

In der Vorlesung haben Sie die Heuristiken Precision, Accuracy, Weighted Relative Accuracy, Gini-Index und ihre äquivalenten Berechnungen kennen gelernt.

Zeigen Sie die Äquivalenz von

- WRA $h_{WRA} = \frac{p+n}{P+N}\left(\frac{p}{p+n} - \frac{P}{P+N}\right)$ und $\frac{p}{P} - \frac{n}{N}$
- Gini-Index $h_{Gini} = 1 - \left(\frac{p}{p+n}\right)^2 - \left(\frac{n}{p+n}\right)^2$ und $\frac{pn}{(p+n)^2}$

Aufgabe 3 CN2's likelihood ratio statistics

Signifikanz-Niveau	0,9	0,95	0,975	0,99	0,995
Schwellen-Wert	2,71	3,84	5,02	6,64	7,88

Gegeben sei ein Datensatz, der aus 60 positiven und 40 negativen Beispielen besteht.

a) Berechnen Sie für die folgenden Regeln, von denen Ihnen nur die Abdeckung bekannt ist, die "CN2's likelihood ratio statistics" und bestimmen Sie für die oben gegebenen Signifikanz-Niveaus, ob die Regeln gepruned werden würde. Hinweis: Verwenden sie den natürlichen Logarithmus.

- R1: p=11 und n=3
- R2: p=15 und n=2
- R3: p=22 und n=6

b) Überlegen Sie sich ohne Berechnung der "CN2's likelihood ratio statistics", warum eine Regel, die 9 positive und 6 negative Beispiele abdeckt, für alle Signifikanz-Niveaus gepruned wird.