Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Matthias Schubert
Erich Schubert

Knowledge Discovery in Databases II

WS 2013/2014

Übungsblatt 12: Stream Data Mining

Aufgabe 12-1 Hoeffding trees

Predict the risk class of a car driver based on the following attributes:

- Time since getting the driving license ($1-2$ years, $2-7$ years, >7 years)
- Gender (male, female)
- Residential area (urban, rural)

These are the first 8 examples.

Person	Time since license	Gender	Area	Risk class
1	$1-2$	m	urban	low
2	$2-7$	m	rural	high
3	>7	f	rural	low
4	$1-2$	f	rural	high
5	>7	m	rural	high
6	$1-2$	m	rural	high
7	$2-7$	f	urban	low
8	$2-7$	m	urban	low

- Incrementally construct a Hoeffding tree for this example.
  Use information gain and $\delta = 0.2$ and $N_{\text{min}} = 2$.
- Compute the value of δ at which the tree would still consist of the leaf only.

Aufgabe 12-2 Cohen's Kappa

Gegeben seien die folgenden Konfusionsmatrizen zu den Zeitpunkten $t = 1, 2, 3$:

$t = 1$

	positiv	negativ
positiv	37	14
negativ	17	32

$t = 2$

	positiv	negativ
positiv	65	8
negativ	7	20

$t = 3$

	positiv	negativ
positiv	90	4
negativ	5	1

Berechnen Sie Accuracy und Cohen's Kappa, und vergleichen Sie die Ergebnisse.

Aufgabe 12-3 k-means auf Streams

Von k-means sind zwei Varianten geläufig:

- Lloyd-Forgy-Variante: 2 Phasen, Objekte zuweisen und Mittelwerte aktualisieren
- MacQueen-Variante: immer je 1 Objekt zuweisen, Mittelwert aktualisieren

Rufen Sie sich beide Varianten in Erinnerung. Auf Stream-Daten kann MacQueen unverändert verwendet werden (kann aber mit "concept drift" nicht gut umgehen). In verteilten Systemen ist Lloyd einfacher umzusetzen.

Überlegen Sie sich, woran man festmachen/erkennen kann, wieso die eine Variante einfacher für Streams einfacher zu adaptieren ist, die andere einfacher für parallele Systeme verwendet werden kann.