JProf. Dr. Uwe Ligges, Prof. Dr. Katharina Morik
Julia Schiffner, Felix Jungermann

Dortmund, 04. Mai
Abgabe: bis zum 11.05. um 10:00 Uhr an
felix.jungermann@tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2010

Blatt 3

Aufgabe 3.1 (7 Punkte)
Betrachten Sie folgenden Datenstrom:

id			
1	1	2	3
2	1	1	1
3	1	2	3
4	1	2	2
5	1	2	1
6	1	1	1
7	1	2	3
8	1	2	1

(a) Wenden Sie den aus der Vorlesung bekannten *Full Ancestry*-Algorithmus an! Benutzen Sie $\epsilon = 0,25$ und $\Phi = 0,25$ (auch wenn normalerweise $\epsilon \ll \Phi$), generalisieren Sie jeweils (wie auch in der Vorlesung) die Suffixe! Geben Sie jeweils die entstehenden Bäume (mit allen Werten wie g, m und Δ) nach Durchlaufen der Fenster und nach den Aufrufen der *Compress*-Methode an.

(b) Welche Elemente werden als *Hierarchical Heavy Hitters* ausgegeben?

(c) Welches Element würde bei der direkten Berechnung der häufigen Mengen auch ausgegeben werden?

Aufgabe 3.2 (3 Punkte)
Für diese Aufgabe benötigen Sie das aus der Übung bekannte Programm *RapidMiner*.

Laden Sie den Prozess *Blatt3.rpm*, der auf den Vorlesungsseiten verfügbar ist, in RapidMiner ein!

(a) Probieren Sie bei dem *Validation*-Operator alle möglichen Parameter für *sampling type* aus (*linear sampling*, *shuffeled sampling* und *stratified sampling*)! Beschreiben und begründen Sie die unterschiedlichen Ergebnisse!

(b) Welches aus der Vorlesung bekannte Sampling-Verfahren fehlt bei den oben angesprochenen Parametern für *sampling type*? Überlegen Sie sich ein Beispiel für das dieses fehlende Sampling-Verfahren nützlich ist!

Geben Sie bei der Bearbeitung aller Aufgaben den Rechenweg mit an!