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

Knowledge Discovery in Databases II

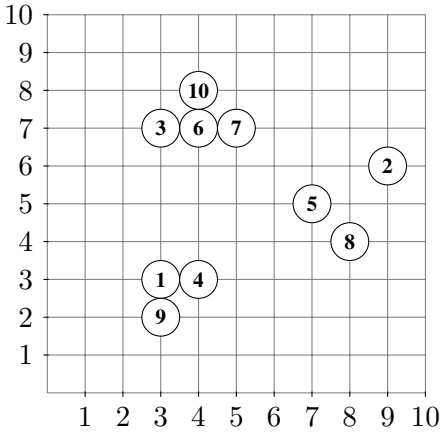
WS 2013/2014

Übungsblatt 10: Microclustering

Aufgabe 10-1 BIRCH Cluster Features

Gegeben folgender Datensatz:

ObjID	X	Y
1	3	3
2	9	6
3	3	7
4	4	3
5	7	5
6	4	7
7	5	7
8	8	4
9	3	2
10	4	8



Berechnen Sie die BIRCH Cluster Features $CF = (N, \vec{LS}, SS)$

Der Radius R eines Cluster Features CF sei definiert als:

$$R(CF) := \sqrt{\frac{SS}{N} - \frac{\vec{LS}^2}{N}}$$

Fügen Sie Elemente genau dann in ein bestehendes Cluster Feature ein, wenn durch das Einfügen der Schwellwert $R \leq T = \sqrt{2}$ nicht überschritten wird. Gibt es mehrere Kandidaten, so wählen Sie den mit dem kleinsten neuen Radius R. Andernfalls erzeugen Sie ein neues CF.

Einen vollständigen CF-Baum brauchen Sie nicht konstruieren, sondern verwenden Sie eine Liste von Blättern CF um den Datensatz zu repräsentieren.

Tipp: prüfen Sie $R^2 \leq T^2$ statt $R \leq T$.

Aufgabe 10-2 Berechnung der Varianz

Betrachten Sie die 1-dimensionalen Datensätze:

$$A = \{0, +1, -1, +2, -2, +3, -3, \ldots, +100, -100\} \quad B = \left\{a_i + 10^{10}\right\} \quad C = \left\{a_i \cdot 10^{-10} + 1\right\}$$

Berechnen Sie die Varianz der drei Datensätze mit:

- Dem naiven Ansatz in zwei Durchläufen: berechnen Sie zunächst den Mittelwert, dann die Varianz über die mittlere quadratische Abweichung.

$$\mu := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mathrm{Var}(X) := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

- Mit Hilfe der Steiner-Verschiebung in nur einem Durchlauf:

$$\mathrm{Var}(X) := \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

- Mit Hilfe des von Knuth und Welfort diskutierten Algorithmus:

```
n = 0
mean = 0.0
sum2 = 0.0
for x in data:
        n = n + 1
        delta = x - mean
        mean = mean + delta/n
        sum2 = sum2 + delta*(x - mean)
var = sum2 / (n-1)
```

- Probieren Sie es in ihrer Lieblings-Anwendung aus (e.g. NumPy, R, ELKI).
  Liefert das Programm den korrekten Wert? Verwenden Sie $n - 1$ Freiheitsgrade!
- Zeigen Sie, dass alle drei Methoden mathematisch äquivalent sind.
  (Tipps: kürzen Sie den Term $\frac{1}{n-1}$ frühzeitig. Für die dritte Methode: verwenden Sie vollständige Induktion und bestimmen Sie geeignete Invarianten!)

Verwenden Sie `double` Genauigkeit für ihre Brechnungen.
Was beobachten Sie? Wie können Sie dieses Problem erklären?