

DataMining CheatSheet

Julian Schubert

8. Juli 2021

1 Gütemaße

1.1 Davies-Bouldin Index (DB)

Güte innerhalb des Clusters C_i	$S_i \sqrt{\frac{1}{ C_i } \sum_{x \in C_i} \text{dist}(x, \mu_i)^q}$
Güte Trennung C_i und C_j	$M_{i,j} = \text{dist}(\mu_i, \mu_j)$
$R_{i,j}$ für $i \neq j$	$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$
Davis-Bouldin Index	$DB = \frac{1}{k} \sum_{i=1}^k D_i$ mit $D_i = \max_{i \neq j} R_{i,j}$

2 Distanzfunktionen

2.1 Distanzfunktionen für Cluster

Single Link	$\text{dist} - \text{sl}(X, Y) = \min_{x \in X, y \in Y} \text{dist}(x, y)$
Complete Link	$\text{dist} - \text{cl}(X, Y) = \max_{x \in X, y \in Y} \text{dist}(x, y)$
Average Link	$\text{dist} - \text{al}(X, Y) = \frac{1}{ X \cdot Y } \cdot \sum_{x \in X, y \in Y} \text{dist}(x, y)$

3 OPTICS

Beschreibung in Worten:

1. Über alle Punkte iterieren
2. Wenn Punkte im Umkreis vom aktuellen Punkt liegen Distanzen updaten
3. Alle Nachbarn vom Punkt abarbeiten
4. Sortiert in die Liste einfügen

4 Assoziationsregeln

- **Support:** $\text{supp}_D(X) = \frac{|\{T \in D \mid X \subseteq T\}|}{|D|}$
- **Frequency:** $\text{supp}_X(D) \cdot |D|$

- **Confidence:** $conf_D(X \rightarrow Y) = \frac{supp_D(X \cup Y)}{supp_D(X)}$