DataMining CheatSheet

Julian Schubert

9. Juli 2021

1 Gütemaße

1.1 Davies-Bouldin Index (DB)

Güte innerhalb des Clusters C_i	$S_i \sqrt[q]{\frac{1}{\lvert C_i \rvert} \sum_{x \in C_i} \text{dist}(x, \mu_i)^q}$
Güte Trennung C_i und C_j	$M_{i,j}$ =dist(μ_i, μ_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	dist - sl(X, Y) = $\min_{x \in X, y \in Y}$ dist(x, y)
Complete Link	dist - cl(X, Y) = $\max_{x \in X, y \in Y}$ dist(x, y)
Average Link	dist - al(X, Y) = $\frac{1}{\lvert X \rvert \cdot \lvert Y \rvert} \cdot \sum_{x \in X, y \in Y}$ 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**: $supp_D(X) = \frac{\lvert \{T \in D \mid X \subseteq T\} \rvert}{\lvert D \rvert}$
- **Frequency:** $supp_X(D) \cdot \lvert D \rvert$
- **Confidence:** $conf_D(X \rightarrow Y) = \frac{supp_D(X \cup Y)}{supp_D(X)}$

5 Auswahl von Assoziationsregeln

5.1 Added Value

$$\frac{sup(A \wedge B)}{sup(A)} - sup(B) = conf(A \rightarrow B) - sup(B)$$

Um wie viel steigt die Wahrscheinlichkeit von B, wenn die Bedingung A Hinzugefügt wird?