

DataMining CheatSheet

Julian Schubert

12. 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:** $\text{conf}_D(X \rightarrow Y) = \frac{\text{supp}_D(X \cup Y)}{\text{supp}_D(X)}$

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?

5.2 Kriterien für Interessantheitsmaße

1. Conciseness:
Kürzere Regeln sind besser (weniger Items)
2. Generality:
Generelle Regeln sind besser (mehr Fälle abgedeckt)
3. Reliability:
Hohe confidence / accuracy ist besser
4. Diversity:
Regeln sollten untereinander unähnlich sein
5. Novelty:
Vorher unbekannt, nicht aus anderen Regeln ableitbar
6. Surprisingness / Unexpectedness:
Gute Regeln widersprechen Vorwissen / Erwartungen
7. Applicability:
Kann praktisch (in der Anwendung) umgesetzt werden

5.3 Monotonie

- **Monotonie:** If a set S violates C, its supersets **might not** violate C, while its subsets **must** violate C
- **Anti-Monotonie:** If a set S violates C, its supersets **must** violate C, while its subsets **might not** violate C

6 Hierarchische Assoziationsregeln

- **Definition** Hierarchische Assoziationsregel:
 $X \Rightarrow Y$, mit $X \subseteq I, Y \subseteq I, X \cap Y = \emptyset$
Kein Item in Y ist sVorfahre eines Items in X (bezüglich H)
- **Support s** einer hierarchischen Assoziationsregel $X \Rightarrow Y$ in D:
Support der Menge $X \cup Y$

- **Konfidenz** c einer hierarchischen Assoziationsregel $X \Rightarrow Y$ in D :
Prozentsatz der Transaktionen, die auch die Menge Y unterstützen, in der Teilmenge aller Transaktionen, welche die Menge X unterstützen

7 Gütemaße für Klassifikation

K: Klassifikator, **TR** Trainingsmenge, **TE** Testmenge

- **Klassifikationsgenauigkeit** G_{TE} :
Alles was aus dem Testset richtig klassifiziert wurde
- **Tatsächlicher Klassifikationsfehler** F_{TE} :
Alles was aus dem Testset falsch klassifiziert wurde
- **Beobachteter Klassifikationsfehler** F_{TR} :
Alles was aus dem Trainingsset falsch klassifiziert wurde

8 Precision und Recall

Precision: $\frac{\text{True positive}}{\text{True Positive} + \text{false Positive}}$

Wenn die Klasse vorhergesagt wird, wie sicher ist die Vorhersage

Recall: $\frac{\text{True positive}}{\text{True Positive} + \text{False Negative}}$

Wie oft wird die Klasse c wieder gefunden

9 Gütemaße für Splits

9.1 Informationsgewinn

Entropie:

$$\text{entropie}(T) = - \sum_{i=1}^k p_i \log(p_i)$$

Informationsgewinn:

$$\text{informationsgewinn}(T, A) = \text{entropie}(T) - \sum_{i=1}^m \frac{|T_i|}{|T|} \cdot \text{entropie}(T_i)$$

9.2 Gini-Index

$$\text{gini}(T) = 1 - \sum_{j=1}^k p_j^2$$

Kleiner Gini-Index \Leftrightarrow geringe Unreinheit

Großer Gini-Index \Leftrightarrow hohe Unreinheit

Gini-Index des Attributs A in Bezug auf T ist definiert als

$$\text{gini}_A(T) = \sum_{i=1}^m \frac{|T_i|}{|T|} \cdot \text{gini}(T_i)$$