

1 Grundlagen

- Qualitative Attribute:
 - Variieren nach Beschaffenheit
- Quantitative Attribute:
 - Variieren nach Wert/Zahlen
- Diskrete Attribute:
 - abgestufte Werte
- Stetige Attribute:
 - können im Intervall jeden reellen Wert annehmen

1.1 Skalenniveaus

- Nominal
 - nur Gleichheit oder Andersartigkeit feststellbar (keine Bewertung)
 - stets qualitativ
- Ordinal
 - natürliche oder festzulegende Rangfolge
- Kardinal/Metrisch
 - numerischer Art
 - Ausprägung und Unterschied sind messbar
 - verhältnisskaliert (Absoluter Nullpunkt vorhanden; (Doppelt so viel.))
 - intervallskaliert (Kein Nullpunkt, nur Differenzen)

1.2 Sym. vs asym. Attribute

- Das symmetrische binäre Attribut ist ein Attribut, bei dem jeder Wert gleichwertig ist (w/m)
- Asymmetrisch ist ein Attribut, bei dem die beiden Ausprägungen nicht gleichwertig sind (Testergebnisse oder Vergleich von Umfragen)
- Rauschen Artefakte, Ausreißer

1.3 Datenvorverarbeitung

- Aggregation
- Sampling
- Diskretisierung / Binarisierung
- Transformation
- Dimensionsreduktion
- Feature Subset Selection
- Feature Creation

1.4 Ähnlichkeits- und Distanzmaße

1.4.1 Ähnlichkeit

- Eigenschaften:
- $s(x, y) 0 \leq s \leq 1$
 - $s(x, y) = 0$, wenn $x = y$
 - Symmetry: $s(x, y) = s(y, x)$
- Simple Matching Coefficient (SMC):

$$SMC = \frac{f_{00} + f_{11}}{f_{01} + f_{10} + f_{00} + f_{11}}$$

Jaccard Coefficient:
Extended Jaccard Coefficient (Tanimoto)
Cosine Similarity:
Correlation:

1.4.2 Distanz (Minkowski)

- Eigenschaften:
- Positivity ($d(x,y) \geq 0$, $d(x,y) = 0$, wenn $x = y$)
 - Symmetry ($d(x,y) = d(y,x)$)
 - Triangle Inequality ($d(x,z) \leq d(x,y) + d(y,z)$)

$$d(x, y) = \sqrt[r]{\sum_{k=1}^n |x_k - y_k|^r}$$

Name	r	Anwendung
Hamming	1	Bin.Vekt.
CityBlock	1	nur gerade
Euclid	2	schräg
Supremum	∞	nur größte Dist.

1.4.3 Weiteres

Verhalten für Multiplikation und Addition:

Property	Cosine	Correlation	Minkowski
Invariant to multiplication	Yes	Yes	No
Invariant to addition	No	Yes	No

Mutual Information:

- Ähnlich wie Correlation, aber für nicht linearen Zusammenhang
 - 0 = kein Zusammenhang, 1 = starker Zusammenhang
- Umrechnung Ähnlichkeit < – > Distanz

2 Klassifikation

3 Clustering

Übungsaufgaben und Musterlösungen