Ludwig-Maximilians-Universität München München, den 8.7.2009
Institut für Informatik
Priv. Doz. Dr. Matthias Schubert, Dr. Arthur Zimek
Erich Schubert

Knowledge Discovery in Databases II

SoSe 2009

Übungsblatt 9: Multi-Instanz Data Mining

Besprechung am 9.7.2009

Aufgabe 9-1 *Distanzmaße für Multi-Instanz-Objekte*

In der Vorlesung wurden diverse Distanzmaße für Multi-Instanz-Objekte vorgestellt. Ein Multi-Instanz-Objekt O_i ist dabei eine Menge von Objekten o_i aus einem Repräsentationsraum R, d.h., $O_i \subseteq R$.

Für ein Distanzmaß $dist : R \times R \to \mathbb{R}_0^+$ sind Distanzmaße für Multi-Instanz-Objekte definiert wie folgt:

Hausdorff:

$$d_{Hausdorff}(O_1, O_2) = \max\left(\max_{o_i \in O_1}\left(\min_{o_j \in O_2}\left(dist(o_i, o_j)\right)\right), \max_{o_i \in O_2}\left(\min_{o_j \in O_1}\left(dist(o_i, o_j)\right)\right)\right)$$

Minimal Hausdorff:

$$d_{MinimalHausdorff}(O_1, O_2) = \min_{o_i \in O_1}\left(\min_{o_j \in O_2}\left(dist(o_i, o_j)\right)\right)$$

Sum of Minimal Distances:

$$d_{SMD}(O_1, O_2) = \frac{1}{2}\left(\frac{1}{|O_1|}\sum_{o_i \in O_1}\left(\min_{o_j \in O_2}\left(dist(o_i, o_j)\right)\right) + \frac{1}{|O_2|}\sum_{o_j \in O_2}\left(\min_{o_i \in O_1}\left(dist(o_i, o_j)\right)\right)\right)$$

Minimal Matching Distanz – o.B.d.A. sei $|O_1| \geq |O_2|$,
$\Pi(O_1)$ sei die Menge aller Permutationen der Instanzen von O_1,
$w(o_{i,j})$ sei ein Straf-Faktor für ungematchte Instanzen:

$$d_{MM} = \min_{\pi_i \in \Pi(O_1)}\left(\sum_{k=1}^{|O_2|} dist\left(O_{1,\pi(k)}, O_{2,k}\right) + \sum_{l=|O_2|+1}^{|O_1|} w\left(O_{1,\pi(l)}\right)\right)$$

Wägen Sie Vor- und Nachteile dieser Distanzmaße gegeneinander ab. Betrachten Sie dazu insbesondere, ob folgende Eigenschaften jeweils gelten, die in ihrer Gesamtheit eine Metrik definieren:

Für ein Distanzmaß $dist : S \times S \to \mathbb{R}_0^+$ und beliebige Objekte $x, y, z \in S$ gilt:

(a) $dist$ ist reflexiv, wenn: $x = y \Rightarrow dist(x, y) = 0$

(b) $dist$ ist symmetrisch, wenn: $dist(x, y) = dist(y, x)$

(c) $dist$ ist strikt, wenn: $dist(x, y) = 0 \Rightarrow x = y$

(d) $dist$ erfüllt die Dreiecksungleichung, wenn: $dist(x, z) \leq dist(x, y) + dist(y, z)$