

Knowledge Discovery in Databases II
 WS 2013/2014

Übungsblatt 7: Multi-Repräsentiertes Clustering, Multi-Instanz-Objekte

Aufgabe 7-1 Multirepräsentiertes Clustering

Gegeben sei ein Datensatz X , so dass jeder Punkt durch 2 zweidimensionale Vektoren repräsentiert wird.

$$\begin{aligned} A &= (0, 1); (3, 0) & B &= (-1, -1); (2, 0) & C &= (0, 0); (3, 1) \\ D &= (0, -3); (-2, 2) & E &= (2, 1); (-2, -3) \end{aligned}$$

Wir wollen auf diesem Datensatz multirepräsentiertes Clustering mittels DBSCAN durchführen.

- Wie unterscheidet sich multirepräsentiertes Clustering von gewöhnlichem Clustering? Welche besonderen Schwierigkeiten sind damit verbunden?
- Es sei $MinPts = 3$. Für welche Werte von $\varepsilon_1, \varepsilon_2$ sind die Objekte C und D Kernobjekte nach
 - der Vereinigungsmethode?
 - der Schnittmethode?

Aufgabe 7-2 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 $dist : R \times R \rightarrow \mathbb{R}_0^+$ sind Distanzmaße für Multi-Instanz-Objekte definiert wie folgt:

Hausdorff:

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

Minimal Hausdorff:

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

Sum of Minimal Distances:

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

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:

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

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 $dist : S \times S \rightarrow \mathbb{R}_0^+$ 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: $dist(x, z) \leq dist(x, y) + dist(y, z)$

Aufgabe 7-3 Hausdorff-Distanz

Zeigen Sie, dass die Hausdorff-Distanz alle drei Eigenschaften einer Metrik erfüllt.

Aufgabe 7-4 Kuhn-Munkres Algorithmus

Gegeben sei folgende Kostenmatrix K. Führen Sie ein Hungarian Matching auf diesen Matrizen durch:

$$K_1 = \begin{bmatrix} 2 & 1 & 3 \\ 6 & 5 & 4 \\ 7 & 8 & 9 \end{bmatrix};$$

$$K_2 = \begin{bmatrix} 90 & 75 & 75 & 80 \\ 35 & 85 & 55 & 65 \\ 125 & 95 & 90 & 105 \\ 45 & 110 & 95 & 115 \end{bmatrix};$$