OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

FAKULTÄT FÜR WIRTSCHAFTSWISSENSCHAFT

Big Data Anwendungen

Aufgabenblatt 4 (Clusteringverfahren)

Aufgabe 1 – Hierachisches Clustering (20 Punkte)

Gegeben seien folgende Daten:

Tag	Anzahl Sonnenstunden	Temperatur	Regen
18	10	-1,0	Ja
17	2	5,0	Nein
25	4	1,0	Ja
30	3	2,0	Ja

(a) Nennen Sie alle Attribute für die mit Hilfe des Jaccard-Index keine Abstände ermittelt werden können und begründen Sie kurz. **(2 Punkte)**

(b) Nennen Sie zwei Verfahren zur Transformation von Daten auf ähnliche Wertebereiche. **(2 Punkte)**

(c) Transformieren Sie die die Attribute „Tag“, „Anzahl Sonnenstunden“ und „Temperatur“ so, dass Sie mit Hilfe der Euklidischen Distanz sinnvoll geclustert werden können. **(3 Punkte)**

(d) Nutzen Sie hierarchisch agglomeratives Clustering zum Clustern der Daten aus (c). **(10 Punkte)**

(e) Zeichnen Sie ein Dendrogramm für die Daten aus (d). **(3 Punkte)**

Aufgabe 2 – kMeans Clustering (18 Punkte)

Gegeben seien folgende Daten:

Höhe	Breite	Farbe	Preis
15	10	rot	50
20	15	gelb	50
17	15	rot	200
13	12	gelb	180
11	16	rot	110
15	12	gelb	120

(a) Geben Sie an welches Attribut Sie für ein kMeans Clustering ohne Transformation nicht nutzen können. Schlagen Sie eine Strategie vor, wie Sie dieses Attribut umwandeln können um es doch zu nutzen. **(3 Punkte)**

(b) Sehen Sie sich die Daten an. Gehen Sie davon aus, dass Sie die Daten in 3 Cluster unterteilen wollen. Geben Sie an, welche Zuordnung der Datenpunkte zu Clustern Sie erwarten, wenn Sie die Euklidische Distanz als Distanzmaß wählen. Erörtern Sie kurz, wie sie zu einem weniger vorhersehbaren Ergebnis gelangen können. **(3 Punkte)**

(c) Wenden Sie den kMeans Algorithmus an um die Daten auf Basis der 3 numerischen Attribute in 3 Cluster zu zerlegen. Nutzen Sie dabei die Euklidische Distanz und lösen Sie das Problem rechnerisch. **(12 Punkte)**

Aufgabe 3 - Verständnisfragen **(10 Punkte)**

(a) Erläutern Sie, wie k Nearest Neighbor Klassifikation funktioniert. Diskutieren Sie dann Vor- und Nachteile des Verfahrens gegenüber anderen Klassifikationsalgorithmen. **(6 Punkte)**

(b) Erläutern Sie kurz was der Unterschied zwischen supervised und unsupervised learning Verfahren ist. **(2 Punkte)**

(c) Erläutern Sie warum die Rangskalierung von Attributen für die Anwendung von Clusteringverfahren hilfreich sein kann und gehen Sie auf Nachteile des Ansatzes ein. **(2 Punkte)**