JProf. Dr. Uwe Ligges, Prof. Dr. Katharina Morik
Julia Schiffner, Felix Jungermann

Dortmund, 13. Juli
Abgabe: bis zum 20.07. um 10:00 Uhr an
felix.jungermann@tu-dortmund.de

Übungen zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2010

Blatt 13

Aufgabe 13.1 (5 Punkte)
Es sind acht Beispiele mit jeweils zwei Attributen gegeben. Die Beispiele und die entsprechenden Attributausprägungen sind in Fig. 1 dargestellt.

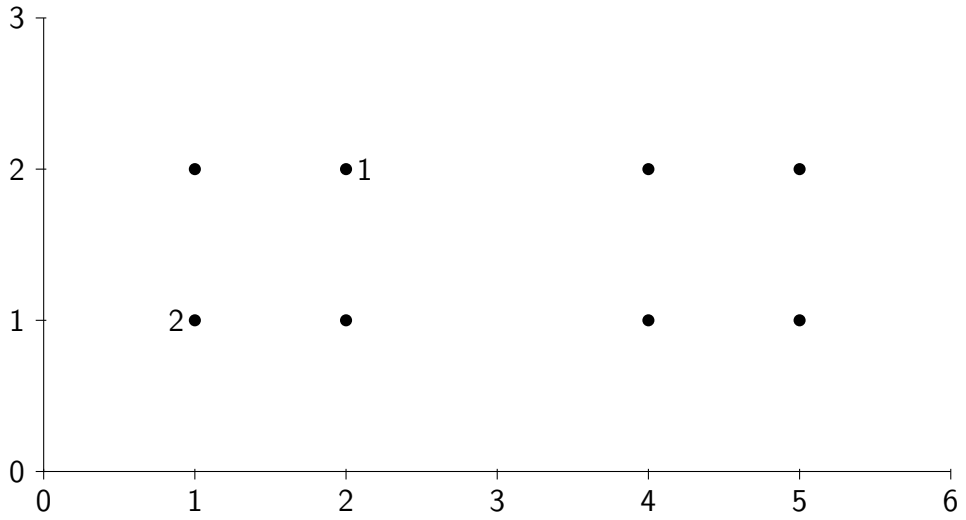

Figure 1: Beispiele mit Attributausprägungen

(a) Führen Sie den K-Means-Algorithmus mit $k = 2$ auf diesen Beispielen händisch aus! Normalerweise werden die Startpunkte für die Mittelpunkte der Cluster (auch Zentroiden genannt) zufällig gewählt. Benutzen Sie jedoch bitte hier die Beispiele $(1, 1)$ und $(2, 2)$ als Startzentroiden! Die Nummerierung neben den Beispielen soll für die Zentroiden beibehalten werden. Falls im laufenden Algorithmus ein konkretes Beispiel äquidistant zu beiden Clusterzentroiden ist, wählen Sie bitte den Clusterzentroid mit kleinerer Nummer als denjenigen Zentroiden aus, dem das Beispiel zugeordnet wird.

(b) In der Vorlesung ist Ineffizienz als ein Nachteil von K-Medoids angegeben worden. Was ist im Vergleich zu K-Means ein weiterer Nachteil? Konstruieren Sie mit Hilfe der Beispiele in Fig. 1 einen Fall, der Ihre Aussage unterstützt!

Aufgabe 13.2 (5 Punkte)
Bei Cluster-Verfahren, deren Clusteranzahl k vom Benutzer vorgegeben werden muss, ist die automatische Bestimmung dieses k kritisch.

(a) Betrachten Sie die Gütefunktion zur Bewertung von Clustern, die aus der Vorlesung bekannt ist:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} D(\vec{x_i} - \vec{x_{i'}})$$

Zur Auswahl welcher Cluster tendiert eine Optimierung, die auf dieser Formel beruht? Was ist somit ein 'optimales' Clustering beruhend auf dieser Formel?

(b) Benutzen Sie nun RapidMiner, um die Iris-Daten mit verschiedenen k-Werten zu clustern. Benutzen Sie den bekannten Operator *Loop Parameter*, um Clusterings für alle ganzzahligen k zwischen 2 und 150 zu erstellen. Benutzen Sie zudem *k-Means*, *Data to Similarity*, *Log* und *Cluster Density Performance* innerhalb der Parameter-Schleife, um die Cluster zu bewerten. *Cluster Density Performance* liefert vergleichbare Ergebnisse wie die Berechnung der oben erwähnten Gütefunktion für Cluster. Lassen Sie sich die Performanz-Werte für die verschiedenen Parameter-Werte k anzeigen und geben Sie diesen Plot zusammen mit der Experiment-Datei ab!

(c) Analysieren Sie den erzeugten Plot und suchen Sie den aus der Vorlesung bekannten "Knick in der Kurve". Gibt es diesen Knick? Was sind u.U. andere Merkmale, die hier ein gutes Clustering auszeichnen. Ihr Wissen über die Beschaffenheit des Iris-Datensatzes ist hier hilfreich.